

32/3

revista do centro de estudos humanísticos  
2018

# diacrítica

corpora nas  
humanidades digitais



**32/3**

revista do centro de estudos humanísticos  
2018

# diacrítica

corpora nas  
humanidades digitais



**Título:** DIACRÍTICA (Nº 32/3 – 2018)  
*Corpora nas Humanidades Digitais*

**Editores:**

Idalete Dias

Sílvia Araújo

Álvaro Iriarte Sanromán

Geoffrey Williams

**Revisão:** Laila Xavier; Orlando Grossegeesse

**Comissão Redatorial:**

Alberto Simões (Instituto Politécnico do Cávado e Ave); Aldina Marques (Univ. do Minho); Anna Cermakova (University of Birmingham); Carlos Garrido (Univ. de Vigo); Cristina Becker Lopes Perna (Pontifícia Universidade Católica do Rio Grande do Sul); Cristina Fargetti (Univ. Estadual Paulista – Júlio de Mesquita Filho, UNESP-Araraquara); Cristina Martins (Univ. de Coimbra); Diana Santos (Univ. de Oslo); Fátima Oliveira (Universidade do Porto); Fernando Alves (Univ. do Minho); Henrique Barroso (Univ. do Minho); Isabel Ermida (Univ. do Minho); Joana Filipa Passos (Univ. do Minho); Joana Vieira Santos (Univ. de Coimbra); José Teixeira (Univ. do Minho); Luiz Antônio da Silva (USP – Univ. de São Paulo); Margarida Pereira (Univ. do Minho); Maria da Conceição de Paiva (Univ. Federal do Rio de Janeiro); Maria José Finatto (Univ. Federal do Rio Grande do Sul, Porto Alegre, UFRGS); Odair Luiz Nadin da Silva (Univ. Estadual Paulista – Júlio de Mesquita Filho, UNESP); Pedro Henriques (Univ. do Minho); Rute Costa (Univ. Nova de Lisboa); Telma São Bento Ferreira (Pontifícia Universidade Católica de São Paulo); Tereza Afonso (Univ. do Minho); Xavier Guinovart (Univ. de Vigo).

**Comissão Científica:**

Abel Barros Baptista (Universidade Nova de Lisboa); Antónia Coutinho (Universidade Nova de Lisboa); António Branco (Universidade de Lisboa); Ana Brito (Universidade do Porto); Augusto Soares da Silva (Universidade Católica Portuguesa); Bernard McGuirk (University of Nottingham); Clara Rocha (Universidade Nova de Lisboa); Conceição Paiva (Universidade Federal do Rio de Janeiro); Eduardo Paiva Raposo (University of California); Fátima Oliveira (Universidade do Porto); Fernando Cabo Aseguinolaza (Universidad de Santiago de Compostela); Graça Rio-Torto (Universidade de Coimbra); Helder Macedo (King's College); Helena Buescu (Universidade de Lisboa); Ivo Castro (Universidade de Lisboa); João de Almeida Flor (Universidade de Lisboa); José Luís Cifuentes Honrubia (Universitat d'Alacant); José Luís Rodrigues (Universidade de Santiago de Compostela); Jürgen M. Meisel (Universität Hamburg / University of Calgary); Maria Alzira Seixo (Universidade de Lisboa); Maria Irene Ramalho (Universidade de Coimbra); Maria João Freitas (Universidade de Lisboa); Maria Manuela Gouveia Delille (Universidade de Coimbra); Mary Kato (Universidade de Campinas); Nancy Armstrong (Brown University); Rui Marques (Universidade de Lisboa); Susan Bassnett (University of Warwick); Susan Stanford Friedman (University of Wisconsin-Madison); Tomás Albaladejo Mayordomo (Universidad Autónoma de Madrid); Vita Fortunati (Università di Bologna); Vítor Aguiar e Silva (Universidade do Minho).

**Edição:** Centro de Estudos Humanísticos da Universidade do Minho em colaboração com Edições Húmus – V.N. Famalicão. E-mail: [humus@humus.com.pt](mailto:humus@humus.com.pt)

Publicação subsidiada por

FCT – Fundação para a Ciência e a Tecnologia

ISSN: 0807-8967

Depósito Legal: 18084/87

Composição e impressão: Papelmunde – V. N. Famalicão

# ÍNDICE

## CORPORA NAS HUMANIDADES DIGITAIS

- 7 **Introdução**
- 11 **Identidade e diferenças na terminologia da *fauna* e da *flora*: notas sobre um estudo comparativo entre as línguas portuguesa, inglesa, italiana e espanhola**  
Sabrina de Cássia Martins
- 31 ***Agroquímico, biocida, pesticida, plaguicida e producto fitosanitario: uma pesquisa com corpus***  
Mauren Thiemy Ito Cereser; Cleci Regina Bevilacqua
- 61 **Quando o léxico dá bandeira – aspectos cognitivo-discursivos da mudança semântica na construção de brasileirismos em registros lexicográficos luso-brasileiros**  
Anderson Salvaterra Magalhães; Janderson Lemos de Souza
- 87 **Características identificadoras e dificuldades na aplicação de listas para a anotação de entidades geográficas mencionadas**  
Afonso Xavier Canosa
- 105 **Uma versão em português europeu do *C-test***  
Masayuki Yamada
- 129 **Aplicação de ferramentas para coleta e análise de dados em linguística**  
Roberlei Alves Bertucci
- 157 **Análise diacrónica dos tempos compostos *tinha feito, terei feito e teria feito* na língua portuguesa**  
Jan Hricsina
- 177 **Análise contrastiva das formas de tratamento ao interlocutor no teatro brasileiro e português dos séculos XIX e XX**  
Ana Carolina Morito Machado
- 207 **The role of pragmatic markers in academic spoken Interlanguage: a corpus-based study of a group of Brazilian EFL university students**  
Bárbara Malveira Orfanó; Ana Larissa Adorno Marciotto Oliveira; Spencer Barbosa da Silva

- 227 **Corpus Stylistics in translation-oriented text analysis:  
Approaching the work of Denton Welch from a Functionalist perspective**  
Guilherme da Silva Braga
- 249 **Em vida e na hora da morte também:  
o que dizem registros de óbito oitocentistas da freguesia de  
Nossa Senhora da Penha de Corumbá (1847–1855)**  
Maria Helena de Paula; Amanda Moreira de Amorim

# Corpora nas Humanidades Digitais

## Corpora in the Digital Humanities



## INTRODUÇÃO

O presente número da revista *Diacrítica* é dedicado aos *Corpora nas Humanidades Digitais*. Na última década, temos assistido a grandes avanços no processamento de dados, motivados pelo estudo e pela análise da imensurável quantidade de informação que se encontra dispersa e cuja produção aumenta de dia para dia. Graças às novas tecnologias e ao desenvolvimento da Ciência da Computação, a Linguística de Corpus abriu caminho ao estudo dos fenómenos linguísticos de forma nunca antes concebida, assim como a outras utilizações desenvolvidas à medida de diversos profissionais (professores, tradutores, linguistas, lexicógrafos, terminólogos, informáticos, historiadores, etc.). No âmbito deste volume, pretende-se dar uma visão das diferentes aplicações da Linguística de Corpus, com especial foco na investigação da linguagem em contexto académico.

Em “Identidade e diferenças na terminologia da *fauna* e da *flora*: notas sobre um estudo comparativo entre as línguas portuguesa, inglesa, italiana e espanhola”, Sabrina de Cássia Martins examina o fenómeno da variação denominativa na terminologia da *fauna* e da *flora* nas línguas portuguesa, inglesa, italiana e espanhola, analisando as variantes denominativas de cerca de duzentas espécies.

Em “*Agroquímico, biocida, pesticida, plaguicida e producto fitosanitário*: uma pesquisa com corpus”, Mauren Thiemy Ito Cereser e Cleci Regina Bevilacqua estudam a equivalência do termo *agrotóxico* em espanhol, num *corpus* de textos legais de países hispanofalantes, a partir dos termos referidos no título do artigo, conforme empregados no cenário de leis ambientais do Brasil e dos países hispânicos.



Em “Quando o léxico dá bandeira – aspectos cognitivo-discursivos da mudança semântica na construção de brasileirismos em registros lexicográficos luso-brasileiros”, Anderson Salvaterra Magalhães e Janderson Lemos de Souza analisam duas unidades simbólicas em que constam tensões conceituais no português do Brasil e no português europeu, *bandeira* e *bandeirante*, em dois trabalhos lexicográficos luso-brasileiros dos séculos XVIII e XIX, que são cotejados com trabalho lexicográfico brasileiro e português do século XXI para fins de identificação, descrição e análise da mudança semântica.

Em “Características identificadoras e dificuldades na aplicação de listas para a anotação de entidades geográficas mencionadas”, Afonso Xavier Canosa descreve as dificuldades na utilização de listas de entidades geográficas (índice de topónimos, ou *gazetteers*) no processo de anotação automática da *Peregrinação* de Mendes Pinto, dado que a simples aplicação dos topónimos da lista pode produzir ambiguidades (*Carvalho*, por exemplo, pode ser um topónimo, um antropónimo ou um nome comum).

Em “Uma versão em português europeu do C-test European-Portuguese version of the C-test”, Masayuki Yamada examina a fiabilidade e a validade de uma versão em português europeu do C-test, um teste de preenchimento simples utilizado para medir a proficiência geral. O teste foi desenvolvido seguindo os procedimentos propostos por Raatz & Klein-Braley (2002). Nele participaram 104 alunos que frequentam cursos de português para estrangeiros em universidades portuguesas.

Em “Aplicação de ferramentas para coleta e análise de dados em linguística”, Roberlei Alves Bertucci mostra que o desenvolvimento crescente de ferramentas informáticas específicas facilita a criação e a análise de *corpora* electrónicos. Neste artigo, são apresentadas três ferramentas (*Netvizz*, *Linguakit* e o *Tropes*) que podem contribuir para o fortalecimento da investigação linguística baseada no uso de *corpora*, e mais especificamente de dados provenientes de redes sociais.

Em “Análise diacrónica dos tempos compostos *tinha feito*, *terei feito* e *teria feito* na língua portuguesa”, Jan Hricsina propõe uma análise diacrónica dos tempos compostos referidos no título do artigo. Efetuada no *corpus* linguístico [www.corpusdoportugues.org](http://www.corpusdoportugues.org), essa análise compara a frequência e o emprego desses tempos na evolução da língua portuguesa.

Em “Análise contrastiva das formas de tratamento ao interlocutor no teatro brasileiro e português dos séculos XIX e XX”, Ana Carolina Morito Machado traça um interessante panorama do sistema de tratamento do interlocutor no PB e no PE dos séculos XIX e XX. As estratégias de referên-

cia ao interlocutor são analisadas em duas amostras de textos dramáticos, à luz do modelo das Tradições Discursivas (TDs), da Teoria da Variação (Weinreich, Labov e Herzog 1968) e da Teoria do Poder e Solidariedade (Brown e Gilman 1960).

No artigo “The role of pragmatic markers in academic spoken interlanguage: a corpus-based study of a group of Brazilian EFL university students”, Bárbara Malveira Orfanò, Ana Larissa Adorno Marciotto Oliveira e Spencer Barbosa da Silva descrevem um estudo que incide sobre o uso de marcadores pragmáticos na produção de discurso oral por parte de estudantes universitários brasileiros inscritos na unidade curricular de Inglês para Fins Acadêmicos. O trabalho baseia-se em dados extraídos do corpus *Brazilian Academic Spoken English* (BRASE) e do subcorpus do *British Academic Spoken English* (BASE).

No texto “Corpus Stylistics in translation-oriented text analysis: Approaching the work of Denton Welch from a Functionalist Perspective”, enquadrando-se no modelo de tradução funcionalista de Christiane Nord, Guilherme da Silva Braga explora a aplicabilidade de uma abordagem baseada em *corpus* numa fase pré-tradutória do texto literário, apontando para as vantagens do tradutor complementar os resultados da análise quantitativa com uma análise qualitativa do texto a traduzir.

O artigo “Em vida e na hora da morte também: o que dizem registros de óbito oitocentistas da Freguesia de Nossa Senhora da Penha de Corumbá (1847–1855)”, da autoria de Maria Helena de Paula e Amanda Moreira de Amorim, aborda importantes aspetos da história e da cultura do período escravocrata do Brasil oitocentista com base nos dados de um *corpus* de registros de óbito da Freguesia de Corumbá de Goiás.

Retomando o tema deste número da *Diacrítica*, é legítimo afirmar que as Humanidades se deparam atualmente com inúmeros desafios tecnológicos que exigem de toda a comunidade científica a necessária abertura a novos paradigmas do conhecimento e da investigação. A construção e a análise de *corpora* nas mais variadas áreas de estudo impulsiona a procura de novos modelos interpretativos que possam ajudar à compreensão de fenómenos culturais, linguísticos, literários e sociais.

*A equipa editorial*



# IDENTIDADE E DIFERENÇAS NA TERMINOLOGIA DA FAUNA E DA FLORA: NOTAS SOBRE UM ESTUDO COMPARATIVO ENTRE AS LÍNGUAS PORTUGUESA, INGLESA, ITALIANA E ESPANHOLA

IDENTITY AND DIFFERENCE IN FAUNA AND FLORA TERMINOLOGY: NOTES ON A COMPARATIVE STUDY AMONG PORTUGUESE, ENGLISH, ITALIAN AND SPANISH LANGUAGES

Sabrina de Cássia Martins\*  
sabrismartins@gmail.com

O presente estudo tem como objetivo examinar o fenômeno da variação denominativa na terminologia da *fauna* e da *flora* nas línguas portuguesa, inglesa, italiana e espanhola. Partindo-se da premissa de que a proximidade entre o observador e o meio em que a espécie ocorre propicia a formação tanto do nome científico quanto de seus nomes vernaculares, analisamos as variantes denominativas em português de cerca de duzentas espécies, comparando as motivações atuantes na formação de tais itens nessa língua, assim como em espanhol, italiano e inglês. Baseamo-nos na Teoria Comunicativa da Terminologia e assinalamos a influência de fatores sócio-históricos e culturais atuantes na composição da terminologia aqui abordada, ênfase para as variantes denominativas compostas por nomes de cores, por nós denominadas de expressões cromáticas especializadas.

**Palavras-chave:** Terminologia da *fauna* e *flora*. Variação. Variação denominativa. Nomes de cores. Expressões cromáticas especializadas.

The present study focus on examining the denominative variation phenomenon in *fauna* and *flora* terminology in Portuguese, English, Italian and Spanish languages. Assuming that the proximity of the observer to the species natural habitat facilitates the formation of both the scientific and vernacular names, we analyse the denominative variants in Portuguese of about two hundred species, comparing the motivation of such items in this language, as well as in Spanish, Italian and English. We based our search on Communicative Theory of Terminology and we affirm that cultural, social and historic factors contribute to the composition of terminology

---

\* Universidade do Estado de Minas Gerais (UEMG), Belo Horizonte, Brasil.

under consideration, with emphasis on denominative variants composed by colour names, *i.e.*, the specialized chromatic phrases.

**Keywords:** *Fauna* and *flora* terminology. Variation. Denominative variation. Colour names. Specialized chromatic phrases.



## 1. Introdução

A virada do século é acompanhada de uma nova concepção sobre o objeto da Terminologia, bem como sobre a contribuição dessa disciplina para a sociedade. Com efeito, fenômenos da linguagem antes refutados por terminólogos passam a ser considerados como fundamentais na formação das unidades lexicais especializadas. Nesse contexto, o presente estudo investiga o fenômeno da variação nos domínios da *fauna* e da *flora* nas línguas portuguesa, inglesa, italiana e espanhola, tomando como base a *Teoria Comunicativa da Terminologia – TCT* (Cabré 1999a; 2003).

Para tanto, utilizamos como referência os nomes populares de cerca de duzentas espécies descritas em Martins (2013a) e que nesse estudo foram divididas em duas subáreas da Biologia: a Botânica, especificamente as *Angiospermas* (monocotiledôneas e eudicotiledôneas), e a Zoologia, exclusivamente os *Vertebrados* (peixes, mamíferos, aves, anfíbios e répteis). Importa explicar que o delineamento do vocabulário em análise é fruto do nosso interesse no uso dos nomes de cores para a ampliação lexical (como já demonstrado em Martins (2013b; 2014; 2017; 2018), em Martins e Zavaglia (2014) e em Zavaglia e Martins (2012; 2016)). A participação no projeto intitulado *Dicionário Multilíngue de Cores (DMC)*<sup>1</sup>, coordenado pela Prof.<sup>a</sup>. Dr.<sup>a</sup>. Claudia Zavaglia, direcionou-nos para o universo terminológico da *fauna* e da *flora*, originando nossas pesquisas de mestrado (Martins 2013a) e de doutorado (Martins 2017). Nesta última, que tem como fruto o presente artigo, analisamos comparativamente a variação denominativa entre as línguas portuguesa (nossa língua materna), inglesa e italiana (nossas línguas de trabalho), observando as motivações atuantes na formação de tais itens, ênfase para a característica cromática das espécies. Às conclusões apresentadas na tese, acrescentamos apontamentos derivados de

1 Para ulteriores informações sobre o projeto, confira Zavaglia (2010; 2007; 2006a; 2006b).

estudos preliminares sobre a língua espanhola, cuja menção explica-se pelo número de países latino-americanos que a têm como língua oficial, inclusive no que se refere ao compartilhamento dos biomas com o Brasil.

Nosso trabalho é impulsionado pela proposição de que a proximidade entre a espécie e o observador propicia a formação tanto do nome científico quanto de suas variantes denominativas. Tal hipótese é comprovada, inclusive, pela frequência de variantes denominativas em língua portuguesa do Brasil para a denominação das espécies presentes nos biomas brasileiros. Nas páginas que seguem, discutimos, primeiramente, o papel do léxico no interior de uma comunidade; em seguida, abordamos a inserção das terminologias como parte integrante do léxico das línguas naturais; posteriormente, relatamos a contribuição do fenômeno da variação para a formação das terminologias; por fim, expomos como se dá a variação denominativa na terminologia em questão, utilizando para tanto alguns exemplos. Em suma, assinalamos a influência de fatores sócio-históricos e culturais atuantes na composição da terminologia aqui abordada, ênfase para as variantes denominativas compostas por nomes de cores, por nós denominadas de expressões cromáticas especializadas.

## **2. O léxico, o indivíduo e a sociedade**

De acordo com Silva (2006), a linguagem é a fronteira fundamental entre os seres humanos, sendo ela constituinte da alteridade. Em outras palavras, é a linguagem que delimita a semelhança e a diferença. Assim, como bem afirma a autora, “os lugares do ‘eu’ e do ‘outro’ não são pontos fixos, mas colagens da linguagem que se encarnam nos corpos, nas palavras e na movimentação de sentidos, como nas leituras que são feitas sobre as relações sociais” (Silva 2006, s.p.). Nesse processo, aprendemos a lidar com o que nos é estranho, como também com o que nos é familiar. É essa habilidade que nos possibilita traduzir em palavras os limites entre a identidade e a alteridade, visto que é por meio dos conceitos que relatamos nossa interpretação do mundo. Ademais, possibilita-nos ainda harmonizar a relação entre o individual e a sociedade que nos rodeia.

Dentro desse contexto, importa definir o papel do léxico que, de acordo com Biderman (2001), pode ser entendido como um verdadeiro patrimônio social transmitido através das gerações e que contribui para a formação da herança cultural de um povo. Uma vez que aborda a nomenclatura de todos os conceitos linguísticos e não-linguísticos, é o léxico o responsável

por expressar linguisticamente a visão que uma comunidade que compartilha de uma dada língua tem do mundo que a circunda, segundo suas percepções, sua consciência, suas convicções e seus interesses. É também o léxico, por meio da cristalização dos conceitos, que possibilita a comunicação e a interação social.

A formação das unidades lexicais que compõe o tesouro vocabular, segundo a autora supracitada, deriva de um processo de categorização que exprime a nossa consciência sobre a realidade que vivenciamos. Durante esse processo, baseamo-nos na reunião de traços que auxiliam na identificação e na classificação das entidades que compõem o mundo real. Trata-se de uma atividade mental, estritamente relacionada ao contexto social, frisamos, em que a percepção, a concepção e a interpretação da realidade são memorizadas pelo falante com base em modelos de estruturas semânticas já existentes na língua e, sobretudo, com base no uso. No tocante a este último, podemos dizer que todas as mudanças transcorridas pela sociedade influenciam não apenas no emprego em que ela faz do léxico, mas sobretudo na sua ampliação. Por essa razão, afirmamos que a evolução do léxico é proporcional à evolução sócio-histórica e cultural da comunidade que o utiliza.

Assim, explica-se a diferença numérica entre as línguas no que tange aos itens lexicais integrantes de campos lexicais como nomes de cores, nomes de animais e de plantas. Uma vez que o sistema linguístico está intimamente relacionado aos costumes e à origem do povo que o utiliza, a abundância ou a carência de itens lexicais na denominação e distinção do espectro cromático, por exemplo, condiz com a necessidade confrontada pelo grupo social na interação diária. Tais pressupostos fundamentam o Relativismo Linguístico, que tem em Wilhelm von Humboldt seu principal propulsor. Para esse estudioso, a linguagem é o órgão criativo do pensamento e as diferenças linguísticas derivam de diferentes visões de mundo. Resumidamente, o sistema linguístico transforma e é transformado pela visão de mundo de uma sociedade, por isso a importância de se observar o significado por trás do uso.

Seguindo essa linha de raciocínio antropológica sobre a linguagem, podemos sustentar que é também por meio da palavra que se faz presente a ideologia. Partindo-se do princípio de que ideologia pode ser definida como um conjunto de conceitos fundamentais compartilhados por uma sociedade, sua relação com a linguagem evidencia-se na medida em que o discurso, ou seja, a palavra em uso, materializa a ideologia. Logo, observamos, num primeiro momento, a estruturação de uma rede conceitual, no

âmbito cognitivo, em que se sustentam os conceitos socialmente compartilhados e que serão, num segundo momento, expressos verbalmente na produção de significações distintas, de acordo com sua manifestação na língua.

Assim como a formação das unidades lexicais do nosso dia-a-dia, o processo de nomeação dos conceitos técnicos e científicos está em harmonia com as transformações individuais e sociais, pois também as unidades lexicais especializadas refletem a forma como aquele que as cria vê, percebe e reflete sobre o mundo ao seu redor, visão essa atrelada ao trajeto científico percorrido pela comunidade em que o indivíduo se insere. Afinal, se as terminologias são de fato componentes do léxico de uma língua, torna-se evidente que refletirão toda a diversidade que caracteriza a linguagem humana, resultado das diferenças em se compreender um mesmo conceito (Cabré 1999). A esse respeito, Cabré (1999) enfatiza a origem compartilhada pelos termos e pelas palavras, afirmando que, acima de tudo, são unidades lexicais, uma forma neutra escolhida pelos estudiosos do léxico de modo a evidenciar que, fora de um contexto, são apenas itens associados a uma dada informação, seja ela gramatical, pragmática ou enciclopédica.

A concepção de que as unidades lexicais especializadas são signos linguísticos e, portanto, componentes do léxico de uma língua natural gera o reconhecimento de que tais unidades estão sujeitas aos mesmos fatores que atuam na formação das palavras que permeiam os discursos cotidianos, incluindo as variações e imposições da cultura de cada povo. Tais fatores implicam, muitas vezes, na construção de estruturas linguísticas diversas para a denominação de um mesmo conceito, as chamadas variantes denominativas, isto é, unidades lexicais cuja criação deriva de elementos geográficos e sociais, diferenças de conceituação, adequação ao nível de língua, dentre outros (Freixa 2002). Por isso, tornam-se responsáveis por harmonizar os diferentes graus de conhecimento dos interlocutores que integram uma dada situação comunicativa. Em seguida, discorreremos sobre o processo de criação das unidades lexicais especializadas e o conceito de variação em terminologia.

### **3. A terminologia e a variação**

Afirmamos anteriormente que o léxico é a representação da realidade de uma dada cultura. É por meio dele que expressamos nossa visão do ambiente em que vivemos. Como frisado por Biderman (2001), o léxico pode ser entendido como um todo composto por diversos subsistemas que



compõem a vasta rede semântica em que são enquadrados os conceitos que delineiam o conhecimento. Efetivamente, o vocabulário técnico-científico atua na composição do tesouro lexical e, da mesma forma que o vocabulário utilizado em situações cotidianas, também as unidades lexicais especializadas estabelecem uma relação intrínseca com a cultura do povo que as utiliza, estando sujeitas às variações e imposições desta.

As concepções anteriores resultam de uma mudança de perspectiva nos estudos terminológicos ocorrida a partir da segunda metade do século passado, sobretudo no decorrer da década de noventa, com o surgimento de novas vertentes teóricas que adotaram um viés descritivo de análise, tais como a *Teoria Comunicativa da Terminologia – TCT* (Cabré 1999; 2003). Emergem questões que giram em torno do papel do contexto e da situação comunicacional em que as unidades lexicais especializadas ocorrem, a postura do sujeito que as utiliza, seu nível de conhecimento e do impacto sócio-histórico e cultural na criação do léxico especializado. Ao contrário do tradicionalmente defendido, a TCT prega que uma unidade lexical especializada é um signo linguístico e, portanto, dotada de uma forma e um conteúdo indissociáveis e que, no plano da expressão, representam um conceito. Dessa forma, assume as mesmas características e submete-se às mesmas condições das unidades lexicais presentes nos discursos do dia-a-dia (Cabré 2008a). Nesse contexto, itens terminológicos e itens lexicais passam a ser entendidos como objetos que compartilham de uma mesma estrutura, similares 1. semanticamente, visto que são dotadas de um significado e, possivelmente, relacionadas a mais de um sentido; 2. funcionalmente, pois pertencem a uma categoria gramatical; e também 3. pragmaticamente, já que permeiam a interação comunicativa (Cabré 2008b). Sua diferença, portanto, relaciona-se aos usuários que fazem uso de tais unidades na comunicação, à situação de uso, à temática que veiculam e ao tipo de discurso em que ocorrem (Cabré 1999).

A fim de demonstrar a interdisciplinaridade que envolve o estudo das terminologias, Cabré (1999; 2003) propõe o *princípio da poliedricidade*, o qual destaca que as unidades lexicais especializadas são ao mesmo tempo unidades linguísticas (pois fazem parte do léxico das línguas naturais e, por isso, são submetidas às mesmas influências que as palavras utilizadas no nosso dia a dia), unidades de conhecimento específico (visto que representam uma categorização da realidade) e unidades de comunicação especializada (já que possibilitam a troca de conhecimento entre especialistas e entre especialistas e leigos, divulgando o conhecimento especializado). Desta sorte, o funcionamento das terminologias só poderá ser adequadamente

explicado se considerarmos seus aspectos linguístico, cognitivo e comunicativo, podendo seu estudo adotar uma perspectiva integradora, isto é, abordando cada um desses aspectos, ou limitar-se a um único aspecto (Cabr  1999).

Uma vez admitida a import ncia de se considerar a situa  o comunicacional em que as unidades lexicais especializadas ocorrem e, por consequ ncia, de se observar o comportamento de tais unidades, bem como as caracter sticas do falante que as utiliza, renuncia-se ao ideal de biunivocidade, que tinha por objetivo a normatiza  o promotora da precis o exigida pela comunica  o especializada. Paralelamente, admite-se a exist ncia da varia  o terminol gica, decorrente do uso natural da l ngua e dos prop sitos da comunica  o. Com efeito, o princ pio da varia  o   um dos elementos que fundamentam a TCT, sendo uma das condi   es inerentes ao estudo das unidades lexicais especializadas. Assim, para o estudo da varia  o no interior dos discursos, a TCT estabelece uma s rie de vari veis, dentre elas: a tem tica, os tipos de interlocutores envolvidos na comunica  o, seu n vel de especializa  o, o grau de formalidade, o prop sito e o tipo de discurso.

Hurtado Albir (2011) frisa a exist ncia de uma grada  o do n vel de especializa  o dos discursos considerados especializados que varia desde o grau m ximo, abrangendo os textos direcionados aos especialistas, at  o grau m nimo, incluindo os textos direcionados ao p blico em geral. Inversamente proporcional ao n vel de especializa  o dos discursos   o n vel de varia  o no interior destes, segundo Cabr  (1999). Para a estudiosa, o n vel m ximo de varia  o encontra-se em discursos de divulga  o cient fica, enquanto que o n vel m nimo situa-se nas terminologias normalizadas por comiss  es cient ficas. Portanto, uma determinada tem tica pode ser abordada em diferentes graus de complexidade e especificidade, em diferentes tipos de texto, tanto pelo discurso especializado, quanto pelo discurso comum.

Segundo Freixa *et al.* (2002), existem dois tipos de varia  o terminol gica: a varia  o localizada no plano das denomina  es, a chamada varia  o denominativa, e a varia  o originada de heterogeneidades no plano do conte do, a chamada varia  o conceitual. Interessa-nos a varia  o denominativa, fen meno definido por Freixa (2002) pela exist ncia de diversas denomina  es para um mesmo conceito; em contrapartida, Bach e Su rez (2002) retratam-no como express  es lingu sticas coexistentes e utilizadas por falantes de n veis de especializa  o diversos para se referirem a um mesmo conceito, estabelecendo uma rela  o de sin nimia em diversos graus. Trata-se de um recurso discursivo que busca evitar a redund ncia

presente tanto na linguagem comum, quanto na especializada e que resulta de usos diferentes de uma unidade lexical especializada por parte de uma comunidade, assim como da sua diversidade social, linguística e geográfica.

O presente estudo tem como foco as diferentes formas variantes que denominam as espécies da *fauna* e da *flora* e que representam variações 1) internas, isto é, cognitivas, pois implicam na forma como o ser humano percebe o espectro, categoriza e utiliza os nomes de cores na linguagem, e 2) externas, isto é, sociais, uma vez que determinam os padrões que influenciarão o emprego dos nomes de cores nas mais variadas situações comunicativas. Nas próximas linhas, discutimos sobre a forma como o fenômeno da variação contribui para a ampliação da terminologia em questão.

#### 4. Metodologia de análise e discussões

Frisamos anteriormente que é no léxico em que as características culturais se mostram mais evidentes. É também por meio dele que os diversos grupos sociais que compartilham de uma mesma língua se distanciam, na medida em que expressa verbalmente a ideologia, as crenças, ou até mesmo os objetivos dos indivíduos. Entretanto, também é o léxico responsável por aproximar os povos, pois permite que os conceitos, muitos deles particulares à determinada cultura, percorram o globo. Assim acontece com a terminologia da *fauna* e da *flora*.

Desde os tempos das grandes navegações, o homem tem se maravilhado com descobertas de novas espécies. Uma diversidade inestimável naquele momento e que até hoje encanta pesquisadores ao redor do mundo pela sua complexidade, pela importância para a sobrevivência do próprio ser humano e pela harmonia com o meio em que habitamos. O deslumbre a cada descoberta vem sendo representado linguisticamente na denominação das espécies, sempre tomando como base a característica mais cristalina e acessível ao nosso aparelho visual. Relatos abordados nas obras de Ferronha *et al.* (1993) e Margarido (2000) apontam para a admiração por parte dos europeus da variedade cromática presente na *fauna* e na *flora* e que, a partir de então, tem sido utilizada para a distinção das espécies. Assim, o homem foi criando diferentes formas linguísticas para denominá-las, isto é, os nomes científicos e as formas vernáculas.

Em referimento à classificação biológica (também chamada de classificação científica ou taxonomia) por meio dos nomes científicos, Amabis e Martho (2001) sublinham que o método binomial tem suas raízes no sistema

elaborado por Carolus Linnaeus no século XVIII, tendo como objetivo o agrupamento dos seres vivos de acordo com suas semelhanças. De forma sucinta, a classificação científica organiza os seres vivos e sumariza o conhecimento que temos deles de forma clara e objetiva a partir da semelhança entre certas estruturas dos mesmos. Contudo, nem todas as características são passíveis de compor o nome científico. Quicke (1996) menciona o caso das cores que, embora possam refletir variação intraespecífica ou fatores ambientais, não são usadas na taxonomia como fator de identificação por retratarem uma característica individual que pode variar de ser vivo para ser vivo. Logo, parte-se como pressuposto que a taxonomia não deve retratar as características individuais, mas da espécie como um todo. Definitivamente, o método proporcionou consolidar os parâmetros da evolução biológica, bem como das relações de parentesco entre as espécies. A padronização e o rigor na descrição representam uma ruptura com a classificação proposta por seus antecessores, estabelecendo por definitivo a sistematicidade na classificação. Além disso, Amabis e Martho (2001) afirmam que tal catalogação facilita a troca de informações e, principalmente, o estudo das espécies.

No tocante às formas vernáculas, trata-se de nomes populares que na maioria das vezes são de criação anterior aos próprios nomes científicos (Garrido 2000). Isso porque muito antes do estudo da espécie o ser humano já convivía e, inclusive, aproveitava-se das suas propriedades. Geralmente, retratam suas características físicas, seus hábitos, sua utilização pelo homem e até mesmo crenças a ela relacionadas. Ademais, cada espécie pode estar relacionada a mais de uma forma vernácula, baseada em uma única ou em diferentes características, originando o fenômeno da variação denominativa. Em Martins (2017), defendemos que tais itens representam as preferências cognitivas dos indivíduos que coabitam o mesmo ambiente em que ocorre a espécie. Com efeito, a cor apresenta-se como um traço fundamental para a distinção das espécies, visto que o nosso aparelho visual envia estímulos ao cérebro que permitem sua rápida identificação.

Nos últimos anos, temos nos dedicado ao estudo da contribuição dos nomes de cores na ampliação do léxico de uma língua, em especial, das terminologias. Observando a presença marcante de tais unidades para a denominação das espécies inseridas nos reinos Animal e Vegetal, propomo-nos a descrevê-las em um dicionário especializado, onomasiológico, e que portanto estivesse em concordância com os pressupostos taxonômicos (Martins 2013a; 2017), e que contemplasse em sua nomenclatura apenas unidades lexicais especializadas formadas por um ou mais nomes de cores inseridos em uma tipologia de nomes de cores que segue os padrões

propostos por Berlin e Kay (1969), Arcaini (1991) e Zavaglia (1996), a saber: *vermelho, verde, azul, amarelo, preto, branco, cinza, marrom, rosa, laranja*, sendo acrescentados os nomes de cor *roxo, violeta e anil*. O acréscimo desses três últimos nomes justifica-se pela alternância na constituição do nome popular da espécie decorrente de distinção de frequências distintas em um dado comprimento de onda. Nesse aspecto, influenciam particularidades do falante, tais como os níveis de escolaridade, o gênero, a idade (atuantes na preferência individual), bem como a comunidade em que se insere o falante e sua região geográfica. Nesse sentido, nosso interesse concentra-se no estudo das variantes denominativas da *fauna* e da *flora* tanto no que diz respeito à sua estrutura formal quanto aos aspectos sociais e cognitivos que atuam na sua formação.

Por fim, o recorte das unidades lexicais consideradas em nosso estudo respeita a ocorrência de nomes de cores na formação do nome popular. Para tal recorte lexical, executamos a princípio uma procura nos dicionários de língua portuguesa Houaiss (2009) e Aurélio (2010), disponíveis em CD-ROM. Com o auxílio das ferramentas de busca, inserimos cada nome de cor para chegarmos aos nomes populares. Posteriormente, validamos o uso de cada um dos itens por meio do *Corpus Web*, seguindo o critério de sua ocorrência aliada ao seu nome científico. Isso feito, executamos a classificação biológica das espécies, restringindo-nos àquelas pertencentes aos subdomínios das Angiospermas, isto é, as monocotiledôneas e as eudicotiledôneas, e dos Vertebrados, ou seja, peixes, anfíbios, répteis, aves e mamíferos. Como dito anteriormente, utilizamos como referência os nomes populares de cerca de duzentas espécies, limitação estabelecida para fins de pesquisa. Entretanto, o inventário total de nomes populares em língua portuguesa vem sendo implementado à medida que damos andamento à pesquisa e inclui atualmente 1200 itens.

No que diz respeito às outras línguas de trabalho, a procura pelos correspondentes foi realizada exclusivamente pela Web, sendo utilizados os buscadores o *Google.com*, *Google.it* e *Google.es*, sempre a partir da inserção do nome científico da espécie. Desse modo, a Web não apenas nos possibilitou encontrar os correspondentes dos nomes populares das espécies consideradas, como também nos permitiu validar o seu uso. Uma vez verificada a existência ou não de correspondentes, a análise comparativa entre as línguas portuguesa (na sua variedade brasileira e, quando oportuno, assinalando particularidades da variedade ibérica), inglesa (na sua variedade americana e oportunamente fazendo menção a particularidades de outros países que a têm como língua oficial) e italiana. Cabe enfatizar que, neste

trabalho, acrescentamos ainda considerações preliminares sobre a língua espanhola. Com efeito, a variação denominativa nessa língua mostra um material linguístico de investigação riquíssimo, pois, além de apresentar características similares à língua portuguesa no que concerne à história de invasão e curiosidade sobre o continente americano, evidencia a variação diatópica, em que é possível visualizar a interação entre espécie e comunidade social a partir de diferentes países.

De um modo geral e sucinto, a investigação das variantes denominativas em sua formação e uso nos mostrou curiosidades, dentre as quais apontamos aqui o que segue:

1. Tendo em vista que a maioria das espécies consideradas neste estudo são nativas da América Latina, os dados obtidos em português e espanhol nos leva à confirmação da hipótese de que a proximidade entre o homem e o meio em que ocorre a espécie contribui para a formação de variantes denominativas que irão compor a terminologia em questão. De fato, encontramos nessas duas línguas um vasto número de variantes denominativas (compostas ou não por nomes de cores) que representam as preferências dos falantes localizados em diferentes pontos do hemisfério sul. Como exemplo, podemos citar *aroeira-branca*, também conhecida no Brasil por *aroeira-brava*, *aroeira-de-capoeira*, *aroeirinha*, *bugreiro*, *aroeira-do-brejo*. Tal espécie conta com oito nomes populares em espanhol (*aruera*, *molle de beber*, *molle de Córdoba*, *falso molle*, *chichita*, *chichita colorada*, *molle dulce/blanco*), porém, com apenas um nome popular em inglês (*wild aroeira*) e um em italiano (*aruera*).
2. Em contrapartida, verificamos a ausência de variantes denominativas em língua italiana para um número relativamente alto de espécies, especificamente, 28 espécies da *flora* e 12 espécies da *fauna*, todas elas nativas da América Latina, o que contribui para a conclusão relatada no item anterior.
3. Essa característica é social e geograficamente variável, visto que comunidades mais próximas ao habitat da espécie apresentam uma tendência maior à escolha do nome de cor na composição da variante denominativa, sobretudo no que diz respeito às espécies da *flora*.
4. Em muitos casos, o elemento cor está presente tanto no nome científico quanto em variantes denominativas, compondo as expressões cromáticas especializadas, por exemplo, *Morus nigra* L. – amora-preta. Tal fato, apesar de ser contraindicado, demonstra a importância da *característica cor* também para a comunidade científica.
5. É comum a tradução literal em língua inglesa do nome popular em língua portuguesa, como em *embaúba-vermelha/red embauba*, *braúna-preta/black brauna*, *angico-vermelho/red angico*; ou ainda, a adaptação do nome científico à língua inglesa, como em *Neoraputia alba* (Nees & Mart.) *Emmerichex*

*Kallunki/arapoca-branca/white neoraputia, Nectandra lanceolata* (Nees & Mart.)/*canela-amarela/lanceolate nectandra*. Tais dados indicam a tentativa de difusão do conhecimento e das descobertas feitas pelos cientistas brasileiros para um nível global por meio da tradução especializada.

6. A análise de diferentes tipologias textuais presentes no *Corpus Web*, enquadradas em níveis de especialização de discurso variáveis, tais como artigos e relatórios científicos, textos jornalísticos, blogs e fóruns da área, apontou para a ampla utilização das expressões cromáticas especializadas, tanto por especialistas quanto por leigos. Concluímos que tais unidades atuam ativamente na difusão do conhecimento e na popularização das espécies da *fauna* e da *flora*.

Por outro lado, tomemos como exemplo espécies que não são endêmicas da América Latina, mas sim que foram trazidas junto com os imigrantes que para cá vieram, tais como a espécie *Lilium candidum* L., originária da Ásia e que é cultivada em diversas regiões do globo. Nesse sentido, ressaltamos que, embora partamos do português do Brasil, incluímos em nosso estudo todo nome popular em língua portuguesa que contenha um nome de cor. Tal espécie apresenta nove variantes denominativas em língua portuguesa, dentre elas, duas são expressões cromáticas especializadas. Em espanhol, esse número se expande para onze variantes e, assim como em português, duas são compostas por nome de cor. Em contrapartida, a língua italiana apresenta cinco variantes, enquanto o inglês tem apenas duas. Em ambas as línguas, apenas um item é formado por nome de cor. Tais informações estão explicitadas no quadro abaixo:

**Quadro 1. Variantes denominativas da espécie *Lilium candidum* L**

	Variantes denominativas			
	português	espanhol	italiano	inglês
<b>Composta por item cor</b>	açucena-branca, lírio-branco	azucena blanca, lírio blanco	giglio bianco	white lily
<b>Não composta por item cor</b>	bordão-de-são-josé cebola-cecém cecém copo-de-leite lírio lírio-dos-poetas lis	lírio, azucena, lírio de san antonio, lis, lily, azucena común, lilio, rosa de juno, vara de san josé	giglio della madonna, giglio di san luigi, giglio di sant'antonio, giglio candido	madonna lily

O quadro nos mostra que há a correspondência em relação ao subdomínio cromático utilizado em todas as línguas abordadas, a saber, *branco* - *blanco* - *bianco* - *white*. Ademais, merece destaque a aproximação de tal planta com a religiosidade e a tradição cultural oral que, embora estejam presente nas quatro línguas, mostra-se mais evidente nas línguas espanhola e italiana, demonstrando a influência da cultura desse povo na criação de itens lexicais especializados. Também nessa língua temos a reunião da cor com a religiosidade expressa pela palavra *candido* que remete à sensação cromática e que simboliza a *pureza*.

Vejamos agora o exemplo abaixo, referente à espécie *Benincasa hispida* (Thunb.).

**Quadro 2. Variantes denominativas da espécie *Benincasa hispida* (Thunb.)**

	Variantes denominativas			
	português	espanhol	italiano	inglês
<b>Composta por item cor</b>	abóbora-branca	calabaza blanca, melon blanco	zucca bianca	White Gourd, White Pumpkin, white gourd melon
<b>Não composta por item cor</b>	abóbora-d'água, caravela	calabaza de la cera calabaza cerosa melon de invierno	Zucca della cera, zucca tamburella, melone di inverno	Wax Gourd, winter melon, winter gourd, ash gourd,

Assim como no exemplo precedente, também essa trepaderia é originária da Ásia, além de ser mundialmente produzida e comercializada. Observamos no quadro 2 a multiplicidade de variantes, entretanto, em maior número em língua inglesa, fruto provável da comercialização do fruto e da proximidade do homem ao ambiente em que é produzida. Enfatizamos ainda a presença do nome de cor e a correspondência cromática em todas as línguas.

Em contrapartida, analisemos exemplos endêmicos da América Latina. A espécie *Nectandra globosa* (Aubl.) Mez, por sua vez, é nativa da região Amazônica, o que contribui para a variedade de nomes em línguas portuguesa e, sobretudo, espanhola, devido à abrangência dessa língua no continente americano. Como descrito a seguir, encontramos nomes comuns cujo uso se restringe a determinados países de língua espanhola. Observemos o quadro abaixo:



**Quadro 3. Exemplos de variantes denominativas para espécie *Nectandra globosa* (Aubl.) Mez**

<b>Variantes denominativas</b>				
	português	espanhol	italiano	inglês
<b>Composta por item cor</b>	Canela-amarela, canela-preta, cedro-preto, loiro-vermelho, louro-vermelho	sigua amarillo, quizará amarillo, moena blanca (Peru), aguacatillo negro (Honduras), jigua blanco (Ecuador), laurel Blanco (Venezuela), moena amarilla (Peru)	-----	white silverballi
<b>Não composta por item cor</b>	siuabale surineia	quizará tostão, sigua, tostão, aguacate de mono, aguacatillo (Costa Rica), aguacate posan, lagarto moena (Peru), laurel (Bolívia), moena (Peru), moena hoja (Peru)	-----	globose nectandra silverballi sweetwood

É interessante notar a multiplicidade de variantes denominativas nessas línguas, principalmente no tocante aos nomes de cores utilizados para a formação desses itens lexicais, uma diversidade geograficamente dependente que permite agrupar os indivíduos de acordo com as características que lhes são peculiares. Em português, temos quatro expressões cromáticas especializadas, com variações tanto no nome que acompanha o item cor, a saber, *canela*, *cedro*, *loiro* e *louro*, quanto no próprio item cor, ou seja, *amarelo*, *preto* e *vermelho*. Em espanhol, temos sete expressões cromáticas especializadas com a mesma variação encontrada em português, entretanto, há a preferência pelos domínios cromáticos *amarelo*, *preto* e *branco*. Com efeito, essa variação deriva, em primeiro lugar, das partes da planta que são tomadas como base para sua distinção, ou seja, suas flores, sua

casca, sua madeira; em segundo, da classe de interlocutores que dão nome à planta, isto é, se estudiosos, comerciantes ou profissionais da carpintaria, ou simples observadores.

Observemos agora um exemplo de variação no domínio da *fauna*:

**Quadro 4. Exemplos de variantes denominativas para espécie *Amazona vinacea* Kuhl, 1820**

Nome científico	Variantes denominativas			
	português	espanhol	Italiano	inglês
<i>Amazona vinacea</i>	papagaio-de-peito-roxo, papagaio-caboclo, curraleiro, coraleiro, jurueba, papagaio-curraleiro, téu-téu, crau-crau, aiurueba, anacã, jurueba, papagaio-de-coleira, paracuã, peito-roxo, quero-quero, xauá	Amazona Vinosa, amazona de pecho vinoso, Loro de pecho rojo, loro vináceo	Amazzone vinata, Amazzone vinacea, Amazonia vinacea, Amazonia vinosa	Vinaceous-breasted Amazon, Vinaceous Amazon Parrot, Vinaceous Parrot, Vinaceous Amazon

Uma primeira consideração diz respeito à diferença numérica de variantes entre as línguas. De fato, a língua portuguesa ressalta-se perante as outras tomadas neste estudo. Também nessa língua, observamos as influências de línguas indígenas, tal como em *anacã* e *paracuã*. As outras línguas, em contrapartida, destacam-se pela tomada do nome científico como base para a formação das variantes.

Por fim, uma última consideração ainda sobre exemplo acima, diz respeito aos nomes de cores utilizados para a formação de variantes em português e espanhol. O português utiliza o item *roxo* que, para a visão humana, localiza-se em um comprimento de onda variável entre 380 e 440 nm.<sup>2</sup>

2 A amplitude de onda visível é medida em nanômetro (nm).

Já o espanhol utiliza o item *rojo*, que corresponde ao item *vermelho* em português, localizado no espectro cromático entre 625 e 740 nm. O que causaria o uso dessas duas línguas por itens gráfica e foneticamente semelhantes? Talvez a evolução do termo roxo, que poderia corresponder a outro sentido, isto é, a outra posição no espectro cromático? O contato entre as línguas e, consequentemente, uma tradução ‘errônea’? Ou, adentrando-nos no nível cognitivo, as diferenças da percepção do espectro cromático que, por conseguinte, derivaria em diferenças na denominação? Independentemente das causas, importa ressaltar que tanto o item *roxo* do português quanto o item *rojo* do espanhol compartilham da mesma origem.

## 5. Considerações finais

Em suma, procuramos, no decorrer do presente estudo, enfatizar que o léxico reflete as mudanças sócio-históricas e culturais transcorridas por uma comunidade. Uma vez reconhecido seu papel de classificar e nomear o mundo à nossa volta, compreende-se que ele também é responsável por definir a nossa identidade. Por outro lado, os discursos dão voz ao léxico, na medida em que possibilita a manifestação das diferenças e a aproximação com o que nos é semelhante.

Também o processo de criação das unidades lexicais especializadas está em harmonia com as transformações individuais e sociais, pois, sendo tais unidades parte do léxico geral de uma língua, fica evidente que também elas estão sujeitas às influências culturais. Portanto, também na formação das terminologias identificamos o fenômeno da variação. Nesse sentido, as variantes denominativas da *fauna* e da *flora* originam-se da necessidade de comunicação entre interlocutores de variados graus de conhecimento. Sem dúvida, representam concretamente a diversidade cultural intrínseca ao léxico.

No que diz respeito especificamente às línguas em estudo, observamos uma diferença numérica significativa. De um modo geral, temos um maior número de variantes para as línguas portuguesa e espanhola. Em contrapartida, a língua italiana apresenta o maior número de casos de ausência de correspondentes tanto para as espécies pertencentes à *fauna* quanto à *flora*. Já no tocante à língua inglesa, fica evidente seu papel de língua franca e de intermediadora na comunicação do conhecimento entre as nações, visto que muitas das denominações populares aparecem traduzidas em textos de divulgação científica.

Por fim, salientamos que o respeito à diferença nos faz compreender a evolução das línguas e nos faz identificar nosso papel na sociedade. Como estudiosos do léxico, nosso trabalho é demonstrar que não se trata de uma evolução apenas linguística, mas da sociedade como um todo.

## Referências

- Amabis, J. M. & Martho, G. R. (2001). *Conceitos de Biologia*. São Paulo: Moderna.
- Arcaini, E. (1991). *Analisi linguistica e traduzione*. Bologna: Patron Editore.
- Bach, C. & Suárez, M. M. (2002). La variación denominativo-conceptual en la traducción científico-técnica: El papel de la reformulación. In J. Chabás *et al.* (Eds.), *Translating Science. Proceedings: 2nd International Conference on Specialized Translation* (pp. 119–127). Barcelona: PPU.
- Berlin, B. & Kay, P. (1969). *Basic Color Terms: Their universality and evolution*. Berkeley/Los Angeles: University of California Press.
- Biderman, M. T. C. (2001). *Teoria Linguística*. (2ª ed.) São Paulo: Martins Fontes.
- Cabré, M. T. (1999). *La Terminologia: Representación y comunicación. Una teoría de base comunicativa y otros artículos*. Barcelona: Universitat Pompeu Fabra.
- Cabré, M. T. (2003). Teorías de la Terminología: De la prescripción a la descripción. In G. Adama & V. Della Valle (Eds.), *Innovazione lessicale e terminologie specialistiche* (pp. 168–188). Florença: Leo S. Olschki. (Serie Lessico Intellettuale Europeo, v. 92).
- Cabré, M. T. (2008a). El principio de poliedricidad: La articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología. *IBÉRICA*, 16(1), 9–36.
- Cabré, M. T. (2008b). De la rigidez a la flexibilidad en la concepción de la terminología: El papel de la lingüística. In M. Dins Casas Gómez & I. Rodríguez-Piñero Alcalá, (Eds.), *X Jornadas de Lingüística* (pp. 89–108). Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.
- Ferronha, A. L., Bettencourt, M. & Loureiro, R. M. (1993). *A Fauna Exótica dos Descobrimentos*. Portugal: Ed. Elo.
- Freixa, J. (2002). *La variació terminològica: anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient* (Tese de Doutoramento, Universitat Pompeu Fabra).
- Freixa, J.; Kostina, I. & Cabré, M. T. (2002). La variación terminológica en las aplicaciones terminográficas. In *Actas del VIII Simposio Iberoamericano de Terminología*. Cartagena de Indias. CD-ROM.
- Garrido, C. (2000). Traducción de los nombres vernáculos ingleses de animales en los textos de divulgación científica. In A. Beeby, D. Ensinger & M. Presal (Eds.), *Investigating Translation* (pp. 251–260). Amsterdão: John Benjamins Publishing Company.

- Hurtado Albir, A. (2011). *Traducción y traductología*. (5ª ed.) Madrid: Cátedra.
- Margarido, A. (2000). *As surpresas da flora nos tempos do descobrimento*. Portugal: Ed. Elo.
- Martins, S. C. (2018). O vocabulário das cores para a ampliação lexical: O caso das unidades fraseológicas e paremiológicas. In C. Zavaglia & A. K. G. Simão (Eds.), *Reflexões, tendências e novos rumos dos estudos fraseoparemiológicos* (pp. 161–180). (1ª ed.) São José do Rio Preto: UNESP/IBILCE.
- Martins, S. C. (2017). *Proposta de uma base de conhecimento multilíngue online de expressões cromáticas da fauna e da flora* (Tese de Doutorado, Universidade Estadual Paulista “Júlio de Mesquita Filho”).
- Martins, S. C. (2014). Cultura, cognição e uso: Aspectos de análise das expressões cromáticas fraseológicas e paremiológicas. *Domínios de Lingu@Gem*, 8(2), 118–138.
- Martins, S. C. (2013a). *Dicionário onomasiológico de expressões cromáticas da fauna e flora* (Diss. Mestrado, Univ. Estadual Paulista “Júlio de Mesquita Filho”).
- Martins, S. C. (2013b). As cores da fauna e da flora: Um dicionário especial composto por cromônimos. *Estudos Linguísticos*, 42(1), 245–256.
- Martins, S. C. & Zavaglia, C. (2014). Léxico e cores: as expressões cromáticas contribuindo para a ampliação lexical. *Revista Trama*, 10(20), 83–96.
- Quicke, D. L. J. (1996). *Principles and Techniques of Contemporary Taxonomy*. 2<sup>nd</sup> ed. London: Blakie Academic Professional.
- Silva, C. M. da. (2006). Metáforas da cultura: Diferença e identidade na leitura da vida social. *Revista Espaço Acadêmico*, 67(6), [s.p.]. Disponível em: <[http://www.espacoacademico.com.br/067/67silva\\_cristina.htm](http://www.espacoacademico.com.br/067/67silva_cristina.htm)>. Consultado em: 10 de Setembro de 2014.
- Zavaglia, C. (2010). Dicionário Multilíngue de Cores: a face eletrônica. In C. Xatara. (Ed.), *Estudos em Lexicologia e Lexicografia Contrastiva*, v. 1, pp. 5–286. (1ª ed.) Curitiba: Honoris Causa.
- Zavaglia, C. (2007). A prática lexicográfica multilíngüe: questões concernentes ao campo das cores. In A. N., Isquendo & I. M. Alves (Eds.), *As Ciências do Léxico: Lexicologia, Lexicografia, Terminologia*, v. III, pp. 209–222. (1ª ed.) Campo Grande, São Paulo: Ed. UFMS; Humanitas.
- Zavaglia, C. (2006a). Dizionario Multilingue di Cromonimi: aspetti metodologici e pratici. GLAT-Bertinoro 2006, Forlì. *Actes de GLAT-BERTINORO 2006*, 1, 1–328. Bretagne: ENST Bretagne.
- Zavaglia, C. (2006b) Dicionário e Cores. *Alfa* (IBILCE/UNESP), 50, 25–41.
- Zavaglia, C. (1996). *Os cromônimos no italiano e no português do Brasil: uma análise comparativa* (Diss. Mestrado, Universidade de São Paulo).
- Zavaglia, C. & Martins, S. C. (2016). Simetrias e assimetrias na representação linguística: o caso das unidades lexicais formadas por nomes de cores. *Revista do GEL*, 13(1), 1130.
- Zavaglia, C. & Martins, S. C. (2012). Dicionários especiais: Uma ponte para divulgação e transmissão dos saberes. *GLAT – GENOVA 2012 – Terminologie: textes, discours et*

*accès aux savoirs spécialisés* (pp. 309–319). Genova: Telecom Bretagne – Università degli Studi di Genova.

## Dicionários

- Ferreira, A. B. H. (2010). *Novo Dicionário Eletrônico Aurélio versão 7.0*. (5ª ed.). Curitiba, Brasil: Positivo Informática LTDA. 1 CD-ROM
- Houaiss, A. (2009). *Dicionário Eletrônico Houaiss da Língua Portuguesa*. Versão 1.0. Rio de Janeiro, Brasil: Editora Objetiva. 1 CD-ROM

[recebido em 9 de maio de 2018 e aceite para publicação em 21 de novembro de 2018]



**AGROQUÍMICO, BIOCIDA, PESTICIDA, PLAGUICIDA E  
PRODUCTO FITOSANITARIO:  
UMA PESQUISA COM CORPUS**  
*AGROQUÍMICO, BIOCIDA, PESTICIDA, PLAGUICIDA AND  
PRODUCTO FITOSANITARIO: A CORPUS-BASED RESEARCH*

Mauren Thiemy Ito Cereser\*  
mauren.cereser@gmail.com

Cleci Regina Bevilacqua\*  
cleci.bevilacqua@ufrgs.br

O objetivo deste trabalho é demonstrar a utilização de *corpora* à luz dos princípios teórico-metodológicos da Linguística de *Corpus* no estabelecimento da equivalência do termo agrotóxico em espanhol. São estudados os termos: *agroquímico*, *biocida*, *pesticida*, *plaguicida*, *producto fitosanitário* e *agrotóxico*, conforme empregados no cenário de leis ambientais do Brasil e dos países hispânicos. Para tanto, foram seguidas as seguintes etapas: a) busca das definições dos termos em dicionários e glossários especializados; b) constituição de *corpora* com textos legais de países hispanofalantes para cada um dos termos; c) coleta de contextos utilizando o *AntConc*; d) busca de traços definitórios; e) elaboração de mapas conceituais; f) identificação dos equivalentes. Fazem parte do quadro teórico desta pesquisa a Teoria Comunicativa da Terminologia (Cabré 1999), a Linguística de *Corpus* (Berber Sardinha 2004), a Equivalência Funcional (Gémar 1998).

**Palavras-chave:** Terminologia. Equivalência. Língua espanhola.

This work aims at demonstrating the use of *corpora* in light of the theoretical and methodological principles of Corpus Linguistics in order to establish the equivalence of the term agrotóxico (Portuguese) in Spanish. The following terms, *agroquímico*, *biocida*, *pesticida*, *plaguicida*, *producto fitosanitário* and *agrotóxico*, had their use analysed in Brazilian and Hispanic countries' environmental law documents and contexts. To do this, the following steps were followed: a) search of term definitions in specialized dictionaries and glossaries; b) compilation of *corpora* with legal texts from Spanish-speaking countries for each one of the terms; c) collection

---

\* Universidade Federal do Rio Grande do Sul, Brasil.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.



of contexts using *AntConc*; d) search for defining characteristics; e) design of conceptual maps; f) identification of equivalents. The Communicative Theory of Terminology (Cabr  1999), Corpus Linguistics (Berber Sardinha 2004), Functional Equivalence (G mar 1998) are part of this research framework.

**Keywords:** Terminology. Equivalence. Spanish Language.



## 1. Introdu  o

Este trabalho fez parte do projeto Combinat rias L xicas Especializadas da linguagem legal, normativa e cient fica (ProjeCom), desenvolvido pelo grupo Termisul, da Universidade Federal do Rio Grande do Sul, Brasil. Entende-se por Combinat rias L xicas Especializadas (CLEs) as unidades sintagm ticas ou oracionais, recorrentes e protot picas de situa  es comunicativas de  reas especializadas, que apresentam certo grau de convencionalidade, condicionado pela l ngua, pela  rea de especialidade e pelo g nero textual no qual ocorrem (Bevilacqua *et al.* 2013).<sup>1</sup> Como exemplos de CLEs temos *adulterar agrot xicos* (em portugu s) e *aprovechamiento de residuos* (em espanhol). O objetivo geral do projeto foi criar uma base de dados multil ngue (portugu s, alem o, espanhol, franc s, ingl s e italiano) e *on-line* de CLEs, utilizando as bases textuais e ferramentas disponibilizadas no Acervo Termisul dirigida a tradutores, produtores e revisores de textos.<sup>2 3</sup>

O objetivo do presente trabalho   demonstrar a utiliza  o de *corpora*   luz dos princ pios te rico-metodol gicos da Lingu stica de *Corpus* no estabelecimento da equival ncia em espanhol do termo *agrot xico*.

O interesse pelo termo *agrot xico* surgiu pelo fato de que, na base de dados do grupo Termisul, estavam como equivalentes desse termo em alguns casos *agroqu mico*, em outros *plagu cida* e em outros *pesticida*. Na tentativa de uniformizar e de identificar o equivalente de *agrot xico*, iniciou-se a busca pelos equivalentes e aos poucos foram aparecendo outros termos que poderiam ser poss veis equivalentes do referido termo. Desta maneira, um

---

1 <http://www.ufrgs.br/termisul/cles/>

2 <http://www.ufrgs.br/termisul/legis.php>

3 <http://www.ufrgs.br/termisul/ferramentas/ferramentas.php>

segundo objetivo deste trabalho foi identificar os termos equivalentes em língua espanhola de *agrotóxico* para busca das CLEs incluídas na base.

A partir do termo em português, os termos a serem estudados em língua espanhola são: *agroquímico*, *plaguicida*, *pesticida*, *biocida*, *producto fitosanitario*.

Para dar conta dos objetivos propostos, foram seguidas as seguintes etapas: a) busca das definições dos termos em dicionários e glossários especializados; b) compilação de *corpora* com textos legais de 14 países hispanofalantes para cada um dos termos; c) coleta de contextos utilizando a ferramenta concordanciador do programa AntConc; d) busca de traços definitórios nos contextos coletados; e) elaboração de mapas conceituais para entender as relações existentes entre os termos com base nos dados analisados; e f) identificação dos equivalentes.<sup>4</sup>

Fazem parte do quadro teórico desta pesquisa princípios da Teoria Comunicativa da Terminologia (Cabré 1999), da Linguística de *Corpus* (Berber Sardinha 2004), da Equivalência Funcional (Gémar 1998). Considerando o referencial teórico seguido, destacamos que o trabalho realizado tem caráter descritivo, seguindo teorias mais recentes da Terminologia. Busca, portanto, descrever os termos encontrados em um *corpus* textual especializado relativo ao Direito Ambiental em espanhol com a finalidade de identificar equivalentes para termos em português brasileiro encontrados em textos de igual temática e gênero.

## 2. Agrotóxico

Na lei brasileira, utilizam-se os termos *agrotóxico*, *agrotóxicos e afins* ou *agrotóxicos, (seus) componentes e afins* para referir-se a:

- a. os produtos e os agentes de processos físicos, químicos ou biológicos, destinados ao uso nos setores de produção, no armazenamento e beneficiamento de produtos agrícolas, nas pastagens, na proteção de florestas, nativas ou implantadas, e de outros ecossistemas e também de ambientes urbanos, hídricos e industriais, cuja finalidade seja alterar a composição da flora ou da fauna, a fim de preservá-las da ação danosa de seres vivos considerados nocivos;
- b. substâncias e produtos, empregados como desfolhantes, dessecantes, estimuladores e inibidores de crescimento [...]. (Lei nº 7.802, 1989)

---

4 <http://www.laurenceanthony.net/>

No *corpus* brasileiro de Direito Ambiental do grupo Termisul, o termo *agrotóxico* tem 402 ocorrências.<sup>5</sup> O termo *biocida* tem 7 ocorrências e *pesticida* 2 ocorrências. Isso demonstra que, em português, *agrotóxico* é o termo mais utilizado no âmbito do Direito Ambiental.

São várias as possibilidades para referir-se à *agrotóxico*, no entanto, escolher uma delas é assumir uma posição. Moragas e Schneider (2003) discutem a utilização desses diversos termos. As indústrias que vendem esses compostos utilizam o termo *defensivo agrícola*, já que protegem as plantações da ação de pragas que poderiam causar prejuízos econômicos. Já na literatura anglo-americana, o termo preferido é o *pesticida* (*pesticides*) que, segundo os autores, exprime a ideia equivocada de combater apenas pestes. Nesse mesmo sentido, *praguicida* seria um termo igualmente limitado, visto que esses compostos também agem em organismos que não são considerados pragas. Eles consideram que o termo mais indicado é *biocida*, pois significa que “mata a vida”, incluindo também organismos que não constituem alvos, que acabam sendo atingidos pela ação desses produtos químicos.

Atualmente, vê-se muito a utilização do termo *agrotóxico*, que significaria substância tóxica de uso agrícola. A utilização desse termo surgiu no movimento ambientalista brasileiro do início da década de 80, com a intenção de dar uma conotação forte e pejorativa a esses produtos, alertando a população sobre seus efeitos prejudiciais (Moragas & Schneider 2003).

Como se pode perceber, a escolha do termo pode mostrar o posicionamento de quem o usou. Por exemplo, um agrônomo utilizando o termo *agrotóxico* estaria assumindo que o produto que ele usa é tóxico e nocivo. O mesmo serve para uma indústria que fabrica esse produto, chamar de *defensivo agrícola* ameniza seus efeitos.

Em espanhol, existe o termo *agrotóxico* assim como em português. No entanto, esse termo não é utilizado em seus textos legais, apenas em textos em que seus autores são contra o uso dessa substância, marcando assim sua opinião.

Para encontrar o equivalente de *agrotóxico* em espanhol no âmbito do Direito Ambiental, foi preciso buscar textos legais e os contextos em que os termos são utilizados. Para tanto, foi importante criar *corpora* de textos legais para obter esses contextos.

---

5 <http://www.ufrgs.br/termisul/ambiental.php>

### 3. Fundamentação Teórica

Considerando o objetivo do estudo realizado, que estabeleceu a interface entre Terminologia e Tradução, apoiada pela pesquisa baseada em *corpus*, foi necessário sustentar-se em uma fundamentação teórica que desse conta dessa interdisciplinaridade. Assim, o quadro teórico desta pesquisa sustenta-se nos princípios da Teoria Comunicativa da Terminologia (Cabré 1999), da Equivalência Funcional (Gémar 1998) e da Linguística de *Corpus* (Berber Sardinha 2004). A seguir, tratamos dessas perspectivas, focando os aspectos de interesse do trabalho.

#### 3.1. Teoria Comunicativa da Terminologia

Para Cabré (2002), a Terminologia é um campo de conhecimento interdisciplinar, ou uma interdisciplina, que deve integrar aspectos cognitivos, linguísticos, semióticos e comunicativos das unidades terminológicas, o que a autora chama de teoria de portas. Essas unidades terminológicas, ou termos, são o objeto da Terminologia; elas transmitem o conhecimento especializado, são produzidas dentro de um discurso especializado e seus significados são resultados de uma negociação entre especialistas.

A Terminologia parte de uma linguagem real para dar conta da denominação especializada, por isso toma os dados da documentação, ou seja, os textos (Cabré 2004). Assim, as unidades terminológicas fazem parte das linguagens de especialidade e aparecem de forma natural nos textos especializados; esses textos são produtos elaborados por especialistas e destinados a informar sobre temas de uma área do saber. O caráter especializado de um texto não se identifica pela restrição do tema tratado, mas sim pelas circunstâncias comunicativas específicas e peculiares em que esses textos são produzidos ou pelas finalidades que cumprem. Desse modo, o trabalho terminológico parte da seleção e da análise da documentação especializada do tema em questão (Cabré 1993).

Cabré (2004) considera a Terminologia representativa da diversidade, pluralidade e multifuncionalidade, adaptada ao meio em que é utilizada e concebida com finalidades específicas. A TCT tem muito a colaborar na reflexão da presente pesquisa, visto que, nessa teoria, leva-se em consideração a observação dos dados terminológicos no discurso natural, ou seja, os textos. Para a autora (2005), no discurso especializado oral e escrito, a terminologia é um recurso expressivo e comunicativo, podendo apresentar

redundância, variação conceitual, variação sinonímica e nem sempre produzindo uma perfeita equivalência entre as línguas.

Como o objetivo do trabalho é estabelecer o equivalente de *agrotóxico* em espanhol, tendo como possíveis equivalentes 5 termos – *agroquímico*, *plaguicida*, *pesticida*, *biocida*, *producto fitosanitario* –, aceita-se o fato de que pode haver mais de um equivalente, sendo eles sinônimos ou, ao menos, equivalentes funcionais.

A Terminologia tem-se ocupado do estudo da sinonímia desde o início de suas reflexões, quando o seu maior interesse era desfazer as ambiguidades na comunicação especializada, alcançando a sonhada univocidade. Atualmente, há uma preocupação maior em tratar e abordar o tema da sinonímia, visto que é um fenômeno presente em todas as línguas naturais (Araújo 2010). Para Araújo (2010), estudar a sinonímia em Terminologia se faz necessário por dois aspectos essenciais. Um deles é que existe uma alta frequência de termos sinonímicos em algumas áreas do saber e, em alguns casos, os sinônimos são encontrados em uma mesma obra, o que pode demonstrar que embora não haja uma aceitação por parte dos especialistas, pelo menos eles estão cientes de que os sinônimos existem. Apesar desse fato concreto, existem afirmações de que a existência da sinonímia seria um empecilho para a exatidão na comunicação especializada; no entanto, a sinonímia não deixa de estar presente. Como um dos autores contrários à existência da sinonímia, temos Wüster (1998), que defende que os sinônimos – ou os termos que têm o mesmo significado – não são desejáveis na terminologia.

Para Cabré (1993), duas unidades formais são sinônimas quando são semanticamente equivalentes, pertencem a uma mesma língua histórica e a mesma variedade formal. A autora acrescenta que as formas sinônimas nem sempre correspondem a padrões de relação de equivalência absoluta. No âmbito da TCT, esse fenômeno também é denominado variação denominativa, distinguindo-se da variação conceitual em que para um termo há mais de uma definição ou conceito.

### 3.2. Equivalência Funcional

Para Gémard (1998), na teoria, traduzir um texto jurídico e traduzir um texto de outra área não se tratam de processos diferentes; independentemente da área, um texto é feito de palavras e termos que carregam conceitos mais ou menos complexos e desenvolvidos, essas palavras são organizadas no discurso de acordo com o idioma, a área de conhecimento e a função

do texto. A particularidade de traduzir textos jurídicos é que, além dos problemas da linguagem, são acrescentados os problemas da norma jurídica e dos conceitos que não coincidem com o outro sistema, de modo que o tradutor seria um mediador entre lei e linguagem.

A noção de equivalência funcional prevê que expressões, em textos do mesmo gênero e temática, em um contexto paralelo, expressam a mesma relação semântica e têm efeito pragmático semelhante no texto de partida e no texto alvo (Gémar 1998). Esse autor tem muito a colaborar com o artigo, visto que trata da tradução dos textos do âmbito jurídico e, portanto, da noção de equivalência.

### 3.3. Linguística de Corpus

Segundo Berber Sardinha (2004), a Linguística de *Corpus* se ocupa da coleta e da exploração de *corpora*, com a finalidade de servirem para a pesquisa de uma língua ou variedade linguística, realizando uma exploração da linguagem mediante evidências empíricas extraídas pelo computador. De acordo com o autor (2004), *corpus* é um artefacto produzido para a pesquisa, coletâneas de textos – escritos ou de transcrições de fala – mantidas em arquivos de computador. O *corpus* deve ser composto de dados autênticos e legíveis pelo computador, ter a finalidade de ser um objeto de estudo linguístico, ter conteúdo criteriosamente escolhido e ser representativo de uma língua ou variedade (Berber Sardinha 2004).

Para a Linguística de *Corpus*, a linguagem é um sistema probabilístico, isso significa que há traços mais frequentes que outros e existe a possibilidade de estabelecer uma relação entre traços mais comuns e menos comuns em determinado contexto. Para tanto, é necessária a observação empírica da frequência do emprego, por diversos usuários e contextos. O *corpus* seria uma fonte de informação, já que registra a linguagem natural em situações reais (Berber Sardinha 2004).

A Linguística de *Corpus* na Terminologia se faz importante visto que os termos devem ser identificados e descritos *in vivo*, nos contextos de uso, ou seja, nos textos especializados (Cabré 2005; Bevilacqua 2013). Para Maciel (2013, p.29) “não se coletam termos ou investigam hipóteses sobre as características de uma linguagem especializada sem contar com acervos textuais informatizados”. Isso significa que sem os *corpora* eletrônicos especializados e as ferramentas computacionais seria muito mais difícil identificar algumas características importantes dos textos para caracterizar termos

e demandaria muito mais tempo nas análises (Bevilacqua 2013). Cabré (2005) aponta que a análise de dados baseada em *corpus* permite dispor de materiais adequadamente selecionados para a descrição de alguns fenômenos – sinonímia, por exemplo –, observar e formular generalizações. Além do mais, muda a maneira de se trabalhar com terminologia, já que deixa de ser um processo manual e cada vez mais incorpora recursos tecnológicos.

Quanto aos critérios para construção de *corpus*, é importante que ele seja composto de textos autênticos, de linguagem natural, escrito por falantes nativos. O conteúdo deve ser escolhido criteriosamente, obedecendo a um conjunto de regras estabelecidas pelas pessoas que o estão criando. Desse modo, o *corpus* coletado corresponderá às características desejadas (Berber Sardinha 2000).

Para esse trabalho, o desejo era construir *subcorpora* de textos legais com a temática agrotóxico em espanhol, a coleta teve que ser guiada por um conjunto de critérios que garantissem, entre outras coisas: a) que todos os textos fossem legais e publicados pelo Governo; b) que a língua espanhola de vários países hispanofalantes fosse representada; c) que houvesse uma quantidade aceitável de documentos para cada *subcorpus*. Outro critério é a representatividade, uma amostra deve ter certa extensão. A representatividade está ligada à questão da probabilidade: existe a possibilidade de estabelecer uma relação entre traços que são mais comuns e menos comuns em certo contexto. A extensão do *corpus* compreende três dimensões: 1) o número de palavras, que é uma medida da representatividade do *corpus*, isso significa que quanto maior o número de palavras maior a chance de o *corpus* apresentar palavras de baixa frequência (que são a maioria das palavras de uma língua); 2) o número de textos, isso permite que o tipo textual, gênero ou registro estejam mais bem representados; 3) o número de gêneros, registros ou tipos textuais (Berber Sardinha 2000), nesse caso, quanto mais textos legais diferentes e de países diferentes representados, maior abrangência da área do Direito Ambiental e do termo agrotóxico.

Costuma-se fazer distinção entre dois tipos de análises do *corpus*: qualitativo e quantitativo. Enquanto na análise qualitativa faz-se uma descrição detalhada e completa de um fenômeno linguístico ou do comportamento de uma palavra (ou grupo de palavras), na análise quantitativa, estabelecem-se índices de frequência aos fenômenos linguísticos observados no *corpus* que podem servir para construir modelos estatísticos, que permitem explicar a evidência encontrada. Esses dois tipos de análise não são excludentes, mas se complementam entre si. A análise qualitativa permite que tanto os fenômenos pouco frequentes quanto os mais frequentes recebam a mesma atenção,

melhorando a quantidade e a qualidade de observações realizadas sobre o *corpus*. A análise quantitativa oferece informações estatisticamente significativas e resultados que podem ser generalizáveis (Pérez Hernández 2002).

Para Berber Sardinha (2004), computar e descrever frequências é uma tarefa típica da Linguística de *Corpus*, bem como a observação dos padrões de uso das palavras do *corpus*. Na presente pesquisa, serão realizadas ambas as análises, tanto qualitativa quanto quantitativa: as frequências serão computadas e descritas e os padrões de uso serão observados. Para tanto, utilizar-se-á o *AntConc*, que é um programa de extração linguística que possui ferramentas, como concordanciador e lista de palavras; ambas as ferramentas são utilizadas nessa pesquisa. O concordanciador mostra uma listagem das ocorrências de um item específico com seus contextos, e a lista de palavras apresenta uma listagem com todas as palavras do *corpus* de acordo com o critério de escolha do usuário (do mais frequente para o menos frequente, do menos frequente para o mais frequente, por ordem alfabética, pelo final da palavra).

#### 4. Metodologia

Primeiro, buscaram-se as definições dos termos *pesticida*, *plaguicida* e *agroquímico* em dicionários e glossários especializados, a fim de sanar, em um primeiro momento, as dúvidas sobre a relação existente entre os termos analisados. Esses 3 termos foram escolhidos já que, como dito anteriormente, eram os termos utilizados como sinônimos na base de dados do grupo Termisul. Foram 7 os glossários consultados, escolhidos por estarem disponíveis na biblioteca do grupo Termisul. Eles tinham em comum o fato de serem especializados em Ecologia e Meio Ambiente. As definições foram registradas e contrastadas, para que fosse possível encontrar alguma provável sinonímia entre os termos (ver 5.1). Em nenhum deles foi encontrada a definição de *agroquímico*. Portanto, a consulta não auxiliou na busca dos equivalentes, já que as definições oferecidas nas diferentes obras apresentavam informações contraditórias e, muitas vezes, insuficientes para sanar as dúvidas encontradas. Foi preciso buscar dados complementares sobre o uso real e significado dos termos em textos da área do Direito Ambiental. Para tanto, constituíram-se *subcorpora* com textos legais de países hispânicos para cada um dos termos. Destacamos, uma vez mais, que a utilização de *corpora* textuais alinha-se à perspectiva proposta pela TCT de identificar e analisar os termos em seus contextos de uso, caracterizando-se como uma perspectiva descritiva. Também segue a perspectiva da equivalência funcional, ou seja, que busca os



equivalentes dos termos em textos comparáveis, tanto em relação ao gênero como à situação comunicativa em que são utilizados e à temática tratada.

Para cada termo a ser analisado, foi compilado um *corpus* específico, totalizando 5 *subcorpora*. Compilar um *corpus* consiste em armazenar todos os textos selecionados, que podem ter sido buscados na internet ou mesmo textos impressos a partir de determinados critérios, como explicitamos a seguir para os *subcorpora* utilizados na pesquisa. A etapa seguinte é preparar o *corpus*, que consiste em converter os formatos para *txt* e fazer a limpeza e formatação (Aluísio & Almeida 2006).

Cada *corpus* foi constituído de textos legais de países hispanofalantes que tinham em seu título ou objetivo os cinco termos estudados. Foram coletados textos legais dos seguintes países: Argentina, Chile, Colômbia, El Salvador, Equador, Espanha, Guatemala, México, Nicarágua, Paraguai, Peru, Porto Rico, República Dominicana, Uruguai. O número de *tokens* (número de palavras) e de *types* (número de palavras desconsiderando repetições) se dividiu da seguinte forma:

Tabela 1. Número de *tokens* e *types* de cada *corpus*

<b>Termo</b>	<b><i>Tokens</i></b>	<b><i>Types</i></b>
<i>Agroquímico</i>	25.055	3.448
<i>Biocida</i>	50.458	4.327
<i>Pesticida</i>	15.649	2.884
<i>Plaguicida</i>	317.078	16.670
<i>Producto fitosanitario</i>	95.322	6.777
<b>Total</b>	<b>503.562</b>	-

É importante destacar que não era a ideia que todos os *subcorpora* tivessem o mesmo tamanho. Quando se trabalha com vários *subcorpora*, espera-se que todos tenham um tamanho semelhante. No entanto, não é o caso para o presente trabalho, o próprio fato de terem tamanhos diferentes pode indicar os termos mais utilizados.

A partir das propriedades propostas por Berber Sardinha (2004), nosso *corpus* se caracteriza por ser um *corpus* de estudo com conteúdo especializado, escrito por falantes nativos, estático e sincrônico. O conteúdo é especializado, visto que são textos legais que tratam da temática relativa a agrotóxico; todos os textos são de modo escrito e de autores nativos do espanhol. O *corpus* é considerado estático, já que os textos foram pré-definidos

e não se acrescentam textos novos. O *corpus* também é considerado diacrônico, pois compreende um período de tempo, mais atual, sendo o texto mais antigo de 1973 e o mais recente de 2015. Assim, foi importante considerar tanto documentos mais recentes quanto documentos mais antigos, dado que muitos textos legais sobre agrotóxico foram escritos no século passado, e, recentemente, essa discussão foi retomada.

Os primeiros *subcorpora* a serem compilados foram os dos termos *agroquímico*, *pesticida* e *plaguicida*, que eram os termos já utilizados na base de dados do grupo Termisul. A necessidade de compilar o *subcorpus* para *biocida* surgiu depois da busca em glossários e dicionários especializados. Já o *subcorpus* de *producto fitosanitario* foi compilado posteriormente, quando o termo apareceu pela primeira vez nos contextos encontrados em outros *subcorpora*.

Os *subcorpora* somam 503.562 *tokens* e 19.982 *types*. Depois de utilizar a *stoplist*: 262.904 *tokens* e 19.804 *types*. A *stoplist* (composta de 185 palavras) auxiliou a excluir as palavras que não eram de interesse para a pesquisa (artigos, preposições, pronomes etc.) e nos permitiu observar as palavras mais frequentes. Uma amostra dessas palavras encontra-se na tabela abaixo:

Tabela 2. 10 palavras mais frequentes

	Ocorrência	Palavra
1	3128	artículo
2	2471	productos
3	2111	registro
4	1760	uso
5	1686	producto
6	1617	aplicación
7	1539	plaguicidas
8	1497	ley
9	1198	caso
10	1197	presente

Observando essas 10 palavras mais frequentes, vemos que fazem parte da área do Direito Ambiental. *Artículo*, *registro*, *ley*, *caso* são termos relacionados à legislação como um todo; já *productos*, *aplicación*, *uso*, e *plaguicidas* são termos relacionados ao uso de *agrotóxico*.

Aplicou-se a *stoplist* em cada *subcorpus* e o novo resultado de *types*, *tokens* e palavras mais frequentes foram:

Tabela 3. Número de *tokens*, *types* e palavras mais frequentes de cada *corpus*

<b>Termo</b>	<b><i>Tokens</i></b>	<b><i>Types</i></b>	<b>Palavras mais frequentes</b>
<i>Agroquímico</i>	13.454	3.309	Artículo, ley, productos, aplicación, registro.
<i>Biocida</i>	26.463	4.182	Biocidas, artículo, productos, biocida, sustancia.
<i>Pesticida</i>	8.203	2.752	Ley, artículo, productos, art., pesticidas.
<i>Plaguicida</i>	164.960	16.500	Registro, artículo, plaguicidas, uso, producto.
<i>Producto fitosanitario</i>	49.824	6.613	Productos, artículo, fitosanitarios, aplicación, producto.
<b>Total</b>	<b>262.904</b>	-	<b>Artículo, productos, registro, uso, producto.</b>

É interessante observar que o termo *artículo* apareceu em todos os *corpora* na primeira ou segunda colocação, isso se deve ao fato de que esse gênero textual – textos legislativos – contém uma estrutura específica que se organiza por artigos, parágrafos, alíneas. Em todos os *subcorpora*, excetuando-se o de *agroquímico*, os termos analisados (*biocidas*, *pesticidas*, *plaguicidas*, *producto fitosanitario*) apareceram entre as palavras mais frequentes. Tal resultado era esperado, pois se tratavam de textos que tinham esses termos como temática. Observa-se, também, uma preferência em utilizar os termos no plural.

Quanto ao tipo de textos legais, temos os seguintes dados: 19 são *leyes*, 8 *reglamentos*, 7 *decretos*, 6 *resoluciones*, 4 *órdenes*, 2 *decretos ley*, 2 *ordenanzas*, 2 *normas oficiales* e 1 *disposición reglamentaria*. É importante ressaltar que cada país tem sua própria hierarquia e sua maneira de organizar a legislação, sendo difícil encontrar equivalentes em português para todos os tipos de documentos existentes nos países de língua espanhola.

Quanto ao número de documentos (figura 1), obteve-se um total de 51: 5 do termo *agroquímico*, 7 de *biocida*, 5 de *pesticida*, 18 de *plaguicida* e 16 de *producto fitosanitario*. Observa-se que a maioria dos documentos refere-se à *plaguicida* e *producto fitosanitario*, o que pode demonstrar os termos mais utilizados nos textos legais.

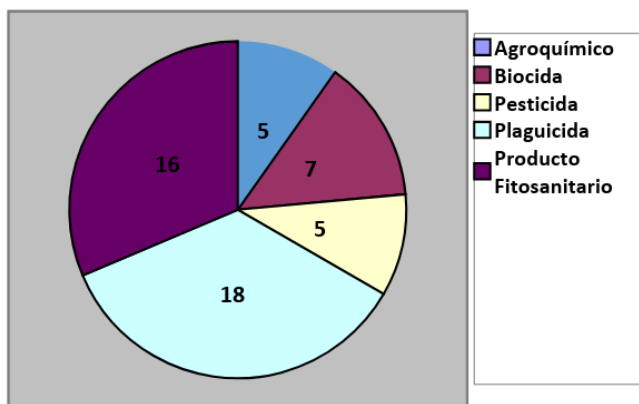


Figura 1. Gráfico da quantidade de documentos para cada termo

Fonte: as autoras

Os países com maior número de textos foram: Argentina com 19 documentos, Espanha com 11 documentos e México com 5 documentos. Isso pode indicar que talvez esses países estejam mais preocupados com essa temática, que eles possuam maior extensão agrícola ou que os outros países não disponibilizam muitos textos legais *on-line*.

Depois de compilar e preparar, os textos devem receber um nome, seguindo um padrão de nomeação (Aluísio & Almeida 2006), o que facilita a recuperação posterior de cada texto. Os textos foram nomeados da seguinte maneira: AGRO para *agroquímico*, BIO para *biocida*, FIT para *producto fitosanitario*, PEST para *pesticida* e PLAG para *plaguicida*. Além de indicar o nome do termo, também foram utilizadas as siglas dos países, para que, quando se analisasse um contexto, pudesse ser observado o país em que tal termo era utilizado. As siglas utilizadas para indicar os países foram: UY, PY, AR, ES, GT, MX, CL, DO, NI, SL, CO, EC, PE, PR.

Para contrastar com esses *subcorpora*, utilizou-se um *corpus* de referência da Real Academia Española, o *Corpus del Español del Siglo XXI* (CORPES XXI), que conta com 237.678 documentos e aproximadamente 225 milhões de formas.<sup>6</sup> Segundo informações do próprio *site*, um *corpus* de referência tem o propósito de servir para a obtenção das características globais que apresenta uma língua em um momento determinado da sua história. Nesse caso específico, o *corpus* deve conter textos de todos os tipos e de todos os países que constituem o mundo hispânico. No CORPES XXI,

6 <http://www.rae.es/recursos/banco-de-datos/corpes-xxi>

também foi utilizada a ferramenta concordanciador, a fim de observar o número de ocorrências e seus contextos.

Com os *subcorpora*, realizou-se a coleta de definições e contextos utilizando a ferramenta concordanciador do programa *AntConc*. Os contextos foram organizados no mesmo documento para compará-los. Nessas definições e contextos, buscaram-se traços definitórios, para que fosse possível identificar as relações entre os termos. Os traços definitórios foram destacados em seus contextos e depois organizados em tabela. Para ilustrar essas relações, elaboraram-se mapas conceituais. Os mapas conceituais são diagramas hierárquicos que explicitam a organização conceitual de uma área especializada, com base nos conhecimentos dos especialistas (Costa 2009). Podem ser utilizados como uma maneira de organizar os termos, permitindo uma melhor visualização das relações entre os termos estudados de uma área do conhecimento específica.

Por fim, depois de analisar a terminologia relacionada a *agrotóxico* em língua espanhola, através de todas as etapas descritas, foi possível identificar os equivalentes para os termos em português.

## 5. Análise dos resultados

A análise dos resultados foi dividida em quatro etapas: análise das definições em glossários especializados, análise dos termos no *corpus* de estudo, *corpus* de referência e elaboração dos mapas conceituais.

### 5.1. Análise das definições em glossários e dicionários especializados

Em um primeiro momento, a fim de entender melhor os termos *pesticida*, *plaguicida* e *agroquímico*, buscaram-se suas definições em glossários e dicionários especializados, com o intuito de identificar se eram sinônimos ou não.

No *Diccionario del medio ambiente* (1984), na entrada de *plaguicida*, não havia nenhuma definição, apenas remetia a *pesticida*. Já na entrada de *pesticida*, foi encontrada a seguinte definição:

pesticida (plaguicida). Agente químico, que suele ser de origen artificial, con el que se elimina a los insectos y otras plagas animales. A veces se ha aplicado como un término general que engloba a los insecticidas, herbicidas, fungicidas, nematocidas, etc. Algunos plaguicidas, como el DDT, han causado efectos sobre muchas especies que no eran objeto de su ataque. (Allaby 1984)

Neste caso, *pesticida* e *plaguicida* são considerados sinônimos, tanto que cada entrada remete a outra. O que também acontece no *Diccionario terminológico de contaminación ambiental* (2000):

Plaguicida (plaguicide). Sinónimo de pesticida. Ver pesticidas. (Martín & Santamaría 2000)

Na definição de *pesticida*, encontrou-se, além de *plaguicida*, outro sinônimo, *biocida*:

Pesticidas (pesticides). Se utilizan como sinónimos los términos plaguicidas y biocidas. Productos químicos, generalmente sintéticos, usados por el hombre para el control y eliminación de diversos organismos vivos, tales como invertebrados (insecticidas, nematocidas, etc.), vertebrados (rodenticidas, avicidas, etc.), plantas (herbicidas), hongos (fungicidas), bacterias (bactericidas), algas (alguicidas), etc. (Martín & Santamaría 2000)

Em duas obras, no *Diccionario Ecológico Ilustrado* (1992) e no *Glosario Ambiental* (1979), há definições diferentes para os dois termos.

Pesticida

Cualquier sustancia o agente utilizado en el control de las pestes; por ejemplo, insecticidas para combatir a los insectos dañinos que atacan a los cultivos agrícolas, fungicidas para el control de las enfermedades producidas por hongos. La fumigación, por diversos medios, ha invadido los campos, dejando tras sí estelas de muerte, cuyas víctimas son ahora los animales.

[...]

Plaguicida

Nombre genérico que se da a los productos que se aplican para atacar los diversos tipos de plagas que afectan a los seres vivos. (Rodríguez 1992)

Neste caso, o *pesticida* seria uma substância enquanto o *plaguicida* seria um produto. No entanto, as funções de "controlar pestes" e "atacar pragas" não parecem ser ações diferentes, o que indicaria uma possível sinonímia; contudo, não há nenhuma remissiva entre as entradas.

PESTICIDA. Cualquier sustancia o agente utilizado en el control de las pestes. Incluye insecticidas, fungicidas, etc.

[...]

PLAGUICIDA. Sustancia química que controla o elimina las plagas. (Medicci & Vivas 1979).

No segundo caso, *plaguicida* seria uma substância química e *pesticida* seria qualquer substância. Levando em consideração essas informações, um *plaguicida* poderia ser um *pesticida*; no entanto, novamente, não há nenhuma remissiva que indique tal relação.

No *Diccionario de ecología, ecologismo y medio ambiente* (Parra 1984), o termo *biocida* aparece novamente em uma definição:

Pesticidas (m.a.) Denominación genérica para aludir a sus sustancias que matan o impiden el crecimiento a ciertos organismos competidores del hombre y sus intereses, sobre todo agrarios. Puede hablarse de herbicidas, o más genéricamente, fitocidas, de fungicidas, alguicidas, bactericidas, insecticidas, etc. Otra palabra empleada es la de biocidas. (Parra 1984)

Nesse dicionário, ignora-se a existência de *plaguicida* e apenas cita-se um sinônimo: *biocida*.

No *Glosario sobre Ecología y Medio Ambiente* (Mariscotti 1993) também acontece o mesmo, apenas há a entrada para *pesticida*:

PESTICIDAS: concepto global para todos los productos químicos que se usan en la agricultura para prevenir el ataque en los cultivos de plagas, animales y vegetales. Se clasifican según su composición química y el sector donde se aplican (insecticidas contra insectos; herbicidas contra malezas; fungicidas contra hongos y defoliantes). (Mariscotti 1993)

Nessa definição, *pesticida* é um produto e uma substância química, características apenas relacionadas a *plaguicida* em exemplos anteriores.

No último dicionário consultado, o *Diccionario Ilustrado de las Ciencias* (Larousse 1988), também não há definição de *plaguicida*:

PESTICIDA. Nombre genérico de las sustancias químicas que se emplean para proteger los cultivos contra sus enemigos vegetales y animales. (Véase FUNGICIDA, HERBICIDA e INSECTICIDA). (Larousse 1988)

Diferentemente das demais definições, nessa há remissivas para tipos de *pesticida*: *fungicida* que combate fungos, *herbicida* que combate ervas daninhas e *insecticida* que combate os insetos.

Após a análise das definições das obras especializadas selecionadas, foi possível chegar a algumas conclusões iniciais: a) aparentemente, *pesticida* e *plaguicida* são sinônimos; b) há um terceiro sinônimo, *biocida*; c) nas definições de *pesticida*, há tipos específicos, como *inseticida* e *fungicida*. Logo,

*pesticida* parece ser o termo mais geral; d) há uma preferência pelo termo *pesticida*; e) não foi possível confirmar se *agroquímico* realmente seria um sinônimo de *agrotóxico*. Portanto, a consulta aos dicionários e glossários especializados não respondeu a muitas das dúvidas, já que foi possível verificar que havia contradições entre as informações que apresentavam. Por essa razão, foi preciso identificar e analisar o funcionamento dos termos nos próprios textos da lei, observando como eles são empregados em seus contextos de uso.

## 5.2. Análise dos termos no corpus de estudo

Em primeiro lugar, analisaram-se o número de vezes que aparecia cada termo nos *subcorpora*, a fim de identificar o termo mais utilizado. *Plaguicida* apareceu 2.169 vezes, *producto fitosanitario* 999 vezes, *biocida* 692 vezes, *agroquímico* 234 vezes e *pesticida* 74 vezes. O fato de *pesticida* ter menos ocorrência surpreende, já que parecia ser o termo mais recorrente nos dicionários e glossários especializados.

A ocorrência de cada termo em cada *subcorpus* especificamente também foi analisada, para verificar se os termos coexistem nos mesmos textos. Na tabela abaixo, é possível observar a ocorrência dos termos em cada um deles. Como era de se esperar, em cada *subcorpus*, o termo que serviu de busca para os textos foi o que ocorreu com maior frequência.

Tabela 4. Ocorrência dos termos em cada *subcorpus*

<i>Corpus</i>	<i>Agroquímico</i>	<i>Biocida</i>	<i>Pesticida</i>	<i>Plaguicida</i>	<i>Producto Fitosanitario</i>
<i>Agroquímico</i>	110	-	-	27	1
<i>Biocida</i>	57	645	-	54	9
<i>Pesticida</i>	-	-	69	10	-
<i>Plaguicida</i>	44	39	1	1.978	44
<i>Producto Fitosanitario</i>	26	8	4	96	891
<b>Total</b>	<b>234</b>	<b>692</b>	<b>74</b>	<b>2.169</b>	<b>999</b>

Em seguida, foram buscados contextos definitórios nos *subcorpora*. Uma facilidade é que, geralmente, nos textos legais, há uma seção em que constam as definições de alguns termos. Para *agroquímico*, havia apenas uma definição:



1. a) Agroquímico: sustancia química usada en la agricultura que tiene un efecto plaguicida. (AGRO\_GT)

Fica claro, como já está dito no próprio nome, que agroquímico se usa no contexto agrícola e sua função é exterminar as pragas. Sua definição foi encontrada em um país apenas (Guatemala). A seguir, trazemos as definições de *biocida* nos textos legais:

1. Biocidas: las sustancias activas y preparados que contengan una o más sustancias activas, presentados en la forma en que son suministrados al usuario, destinados a destruir, contrarrestar, neutralizar, impedir la acción o ejercer un control de otro tipo sobre cualquier organismo nocivo por medios químicos o biológicos, de acuerdo con el artículo 2 del Real Decreto 1054/2002, de 11 de octubre. (BIO\_ES\_3)
2. Biocidas: las sustancias activas y preparados que contengan una o más sustancias activas, presentados en la forma en que son suministrados al usuario, destinados a destruir, contrarrestar, neutralizar, impedir la acción o ejercer un control de otro tipo sobre cualquier organismo nocivo por medios químicos o biológicos. (BIO\_ES\_5)

Como se tratam de decretos do mesmo país (Espanha), eles acabam repetindo a definição. O *biocida* é utilizado pelo usuário, ou seja, aquele que adquire o produto em diferentes formas de apresentação, posto que pode ser utilizado para fins distintos. Desse modo, o contexto em que aparecerá será amplo. Por sua vez, as definições de *pesticida* encontradas foram:

1. Pesticida: cualquier producto destinado a ser aplicado en el medio ambiente con el objeto de combatir organismos capaces de producir daños en el hombre, animales, plantas, semillas, y objetos inanimados, con fines sanitarios o domésticos, diferentes a la protección agrícola. (PEST\_CL\_)
2. PESTICIDAS: toda sustancia química o químico-biológica o mezclas de sustancias destinadas a prevenir o combatir plagas o enfermedades en animales y vegetales, tales como: insecticidas, fungicidas, germicidas, nematocidas, acaricidas, moluscocidas, rodenticidas, ornitocidas, bactericidas, viricidas, repelentes, atrayentes y otros productos para uso tanto en los animales como en los vegetales, con la misma finalidad expresada en esta letra: (PEST\_GT)

Na primeira definição, é deixado claro que *pesticida* tem fins sanitários e domésticos, diferentes à proteção agrícola, desse modo, não apareceria em contexto agrícola. A segunda definição tampouco restringe a situação de uso do produto, podendo o termo aparecer em um contexto de meio ambiente em geral. Destaca-se ainda que o termo foi definido em apenas dois países (Chile e Guatemala). O termo *plaguicida* é o que mais tem definições, como vemos a seguir:

1. Para los efectos de esta Ley, plaguicida o producto afín es toda substancia química, orgánica o inorgánica que se utilice sola, combinada o mezclada para prevenir, combatir o destruir, repeler o mitigar insectos, hongos, bacterias, nematodos, Ácaros, moluscos, roedores, malas hierbas o cualquier otra forma de vida que cause perjuicio directo o indirecto a los cultivos agrícolas, productos vegetales o plantas en general. La terminología técnica así como la clasificación que se deba tener de los plaguicidas deberán constar en el correspondiente Reglamento. (PLAG\_EC)
2. Plaguicida: las sustancias o ingredientes activos, así como las formulaciones o preparados que contengan uno o varios de ellos, destinados a cualquiera de los fines siguientes:
  - a) Combatir los agentes nocivos para los vegetales y productos vegetales o prevenir su acción.
  - [...] f) Hacer inofensivos, destruir o prevenir la acción de otros organismos nocivos o indeseables distintos de los que atacan a los vegetales. (PLAG\_ES\_2)
3. XXXVIII. Plaguicida, cualquier sustancia o mezcla de sustancias que se destine a controlar cualquier plaga, incluidos los vectores que transmiten las enfermedades humanas y de animales, las especies no deseadas que causen perjuicio o que interfieran con la producción agropecuaria y forestal, así como las sustancias defoliantes y las desecantes;
  - [...] XLI. Plaguicida de uso agrícola, el plaguicida formulado de uso directo en vegetales que se destina a prevenir, repeler, combatir y destruir los organismos biológicos nocivos a estos;
  - XLII. Plaguicida de uso en jardinería, el plaguicida formulado utilizado en campos de golf y áreas verdes no destinadas al cultivo de productos agrícolas o forestales;
  - [...] XLVI. Plaguicida de uso urbano, el plaguicida formulado para uso exclusivo en áreas urbanas, incluido el usado en predios baldíos y vías de ferrocarril;
  - XLVII. Plaguicida doméstico, el plaguicida formulado que está listo para su aplicación directa en casas habitación y no requiere ser preparado o diluido de ninguna forma;

XLVIII. Plaguicida equivalente, aquel plaguicida técnico, biocida técnico o plaguicida o biocida técnico concentrado que presenta similitud a un perfil de referencia en sus impurezas, en su perfil toxicológico o en ambos, generados por distintos fabricantes; (PLAG\_MX\_1)

4. Plaguicida: Cualquier sustancia o mezcla de sustancias que se destinan a controlar cualquier plaga, incluidos los vectores de enfermedades humanas y de animales, así como las especies no deseadas que causen perjuicio o que interfieran con la producción agropecuaria y forestal, por ejemplo, las que causan daño durante el almacenamiento o transporte de los alimentos u otros bienes materiales, así como las que interfieran con el bienestar del hombre y de los animales. Se incluyen en esta definición las sustancias defoliantes, las desecantes y los coadyuvantes.

3.6 Plaguicida de uso industrial: Plaguicida técnico o formulado utilizado como materia prima en un proceso industrial para la elaboración de plaguicidas o productos de uso directo.

[...] 3.8 Plaguicida de uso agrícola: Plaguicida de uso directo en campo, destinado a prevenir, repeler, combatir y destruir los organismos biológicos nocivos a los vegetales.

[...] 3.11 Plaguicida de uso urbano: Plaguicida formulado que para su aplicación requiere de previo acondicionamiento y es para uso exclusivo de áreas urbanas, por personal autorizado.

3.12 Plaguicida de uso en jardinería: Plaguicida formulado utilizado en áreas verdes no destinadas al cultivo de productos agrícolas. (PLAG\_MX\_3)

5. PLAGUICIDAS: Son todas las sustancias o mezcla de sustancias, destinadas a prevenir, controlar y eliminar cualquier organismo nocivo a la salud humana, animal o vegetal, o de producir alteraciones y/o modificaciones biológicas a las plantas cultivadas, animales domésticos, plantaciones forestales y los componentes del ambiente. Esto incluye sustancias reguladoras del crecimiento, defoliantes, desecantes, agentes alterantes de la fijación de cosechas y sustancias y métodos físicos empleados para preservar los productos agropecuarios, madera y productos de madera; (PLAG\_NI\_1)
6. 'Plaguicida' significa toda sustancia o mezcla de sustancias preparada(s) rotuladas, destinadas, o que tenga la capacidad para contrarrestar, destruir, repeler, prevenir, esterilizar o mitigar la acción de cualquier plaga y cualquier sustancia o mezcla de sustancias preparadas, rotuladas o diseñadas para usarse como defoliador, desecante y regulador de crecimiento. (PLAG\_PR\_)

Na definição 2, o contexto refere-se basicamente à agricultura ou cultivos. Na definição 1, o contexto aponta usos agrícolas; no entanto, como refere-se a 'plantas em geral', pode-se dizer que o termo é utilizado em um

contexto mais amplo do que o agrícola. Nas definições 3 e 4, fica evidente que existem tipos de *plaguicidas*, tendo o termo um uso mais geral; entretanto, um dos tipos é o *plaguicida* de uso agrícola. Na definição 5, fica claro o contexto amplo também e, na definição 6, o contexto em que o produto é utilizado não fica explicitado. Destacamos que esse termo aparece em vários países: Equador, Espanha, México, Nicarágua e Porto Rico. Por fim, trazemos as definições de *producto fitosanitario*:

1. *Producto fitosanitario o Plaguicidas*: Cualquier sustancia o mezcla de sustancias, destinada a prevenir, controlar o destruir cualquier organismo nocivo, incluyendo las especies no deseadas de plantas, animales o microorganismos que causan perjuicio o interferencia negativa en la producción, elaboración o almacenamiento de los vegetales y sus productos. El término incluye desecantes y las sustancias aplicadas a los vegetales antes o después de la cosecha para protegerlos contra el deterioro durante el almacenamiento y transporte. (FIT\_PY\_)
2. *Plaguicida o Producto fitosanitario*: Cualquier sustancia, agente biológico, mezcla de sustancias o de agentes biológicos, destinada a prevenir, controlar o destruir cualquier organismo nocivo, incluyendo las especies no deseadas de plantas, animales o microorganismos que causan perjuicio o interferencia negativa en la producción, elaboración o almacenamiento de los vegetales y sus productos. El término incluye coadyuvantes, fitoreguladores, desecantes y las sustancias aplicadas a los vegetales antes o después de la cosecha para protegerlos contra el deterioro durante el almacenamiento y transporte. (PLAG\_PY)

Não existe uma definição em que o *producto fitosanitario* é definido sozinho, aparentemente, no Paraguai, *plaguicida* e *producto fitosanitario* são sinônimos, pelo uso da conjunção 'o'. O contexto de uso seria agrícola.

A partir da análise das definições nos textos legais, foi possível identificar semelhanças entre os termos estudados quanto à composição e quanto à ação. Quanto à composição, todos são considerados substância ou mistura de substâncias químicas (orgânica ou inorgânica) ou químico-biológicas. Quanto à ação, servem para *destruir, contrarrestar, neutralizar, impedir la acción, prevenir, controlar, combatir, repeler, mitigar, conservar, hacer inofensivos, eliminar*. Quanto às pragas, referem-se a qualquer organismo nocivo, organismos que causam danos ao homem, animais, plantas, sementes, objetos inanimados; qualquer forma de vida que prejudique diretamente ou indiretamente os cultivos agrícolas, produtos vegetais ou plantas em geral; qualquer organismo nocivo à saúde humana, animal ou

vegetal; qualquer organismo nocivo que prejudique a produção, elaboração ou armazenamento de vegetais e seus produtos.

Também foi possível identificar diferenças em relação ao tipo de praga que o produto combate e à sua aplicação, como se pode observar no quadro abaixo:

Termo	Praga	Aplicação
<i>Agroquímico</i>	----	<i>Agricultura</i>
<i>Biocida</i>	<i>Cualquier organismo nocivo</i>	<i>Para el usuario (amplia)</i>
<i>Pesticida</i>	<i>Organismos que producen daños en el hombre, animales, plantas, semillas, y objetos inanimados</i>	<i>Medio ambiente. Fines sanitarios o domésticos, diferentes a protección agrícola</i>
<i>Plaguicida</i> <sup>1</sup>	<i>Cualquier forma de vida que cause perjuicio directo o indirecto a los cultivos agrícolas, productos vegetales o plantas en general</i>	<i>Agricultura</i>
<i>Plaguicida</i> <sup>2</sup>	<i>Cualquier organismo nocivo a la salud humana, animal o vegetal</i>	<i>Amplia</i>
<i>Producto fitosanitario</i>	<i>Cualquier organismo nocivo que perjudica la producción, elaboración o almacenamiento de vegetales y sus productos</i>	<i>Agricultura</i>

Figura 2. Quadro das diferenças de uso e aplicação dos diferentes produtos

Fonte: as autoras

Parece haver uma coerência entre a aplicação da substância em si e a utilização do termo. Desse modo, a substância *agroquímico*, por exemplo, é aplicada na agricultura, e a utilização do termo nos textos especializados também ocorre em contextos agrícolas.

Sobre o termo *agroquímico*, conclui-se que é um termo que ocorre unicamente em contexto agrícola. Tem 22 ocorrências ao lado *plaguicida* e 21 ocorrências ao lado de *biocida*; no entanto, nos contextos, não fica evidente a relação de sinonímia ou não entre os termos.

Sobre o termo *biocida*, em alguns contextos são apresentados os tipos de biocidas, o que incluem também usos não agrícolas como desinfetantes, conservantes, repelentes, biocidas para a higiene humana e veterinária. Tal fato demonstra que seu uso pode ser mais amplo. Este termo ocorre 3 vezes nos *subcorpora* como sendo *plaguicida no agrícola*:

Este Registro hará posible que la gestión de las empresas que trabajan con productos fitosanitarios se separe de la de aquellas cuya actividad se realiza con productos plaguicidas no agrícolas o biocidas. (BIO\_ES\_1)

Esses 3 contextos podem indicar que *biocida* é utilizado em situações mais gerais, diferentes ao agrícola.

Sobre o termo *pesticida*, decidiu-se desconsiderá-lo devido à sua pouca ocorrência no *corpus*. Outro termo que se desconsiderou foi o *producto fitosanitário*; nesse caso, não se deu por falta de ocorrências, mas sim, por contextos que se contradizem. Nas definições das leis aparecia como termo utilizado em contextos agrícolas, no entanto, em outros contextos, o termo parecia ser usado como termo mais abrangente:

Los organismos administradores del seguro de la ley N° 16.744, deberán informar a sus empresas afiliadas sobre los riesgos asociados al uso de pesticidas, plaguicidas y, en general, de productos fitosanitarios. (FIT\_AR\_4)

Como é possível perceber nesse contexto específico, o termo em questão parece que é o que abrange mais termos. Pela necessidade de uma análise mais aprofundada, ainda não é possível chegar a uma conclusão quanto ao emprego do termo.

Por fim, sobre o termo *plaguicida*, há duas possibilidades: utilizar o termo em um contexto mais geral ou utilizar o termo em contextos específicos agrícolas, ou seja, o termo é o mais abrangente, mas é possível utilizar em contexto agrícola sem necessidade de especificar que é usado nesse tipo de contexto. Caso se anseie uma especificação, pode ser utilizado o termo *plaguicida de uso agrícola*, que tem 22 ocorrências nos *subcorpora*.

## 2.1. Corpus de referência

O termo *agrotóxico* pôde ser evidenciado em 168 casos em 28 documentos. É interessante observar que o termo ocorre majoritariamente em países que fazem fronteira com o Brasil: Uruguai (80 ocorrências), Paraguai (59 ocorrências) e Argentina (17 ocorrências). Sendo que em um dos contextos, aparece a influência dos brasileiros, em uma notícia do Paraguai:

Según los lugareños, los brasileños compraron varios terrenos en la zona y utilizan agrotóxico que no solo afectan a los yuyos, sino que se expande en toda la colonia, de acuerdo al viento que está soplando.

Uma hipótese é que poderia ser influência do português; no entanto, não há como confirmar essa hipótese no momento. Outra particularidade é que os termos, geralmente, são utilizados em contextos pejorativos:

1. Dicen que el agro representa el 45% de las exportaciones paraguayas, pero también representa un gran porcentaje de envenenamiento, cáncer y leucemia, debido a los agrotóxicos. ¿De qué sirve ingresar dólares, si los volveremos a gastar en quimioterapia y remedios de sobrevivencia?
2. A su vez, diversas son las formas de destrucción de los recursos naturales asociadas al uso de biocidas; estos agrotóxicos altamente nocivos son en realidad armas químicas que se originaron en las dos guerras mundiales.

Esses dois contextos, o primeiro de uma reportagem paraguaia e o segundo de um texto acadêmico também do Paraguai, corroboram com o que já afirmamos anteriormente, ou seja, que utilizar *agrotóxico* é marcar uma posição, é assumir que é uma substância prejudicial ao meio ambiente e à saúde humana; sendo assim, quando o autor adota esse termo, mostra uma posição contrária ao seu uso. Do mesmo modo, a presença do termo fora do contexto legislativo mostra que *agrotóxico* pode ser um termo em outra área de especialidade, por exemplo, na agricultura e na saúde. Porém o termo não é utilizado em textos da área pretendida: no Direito Ambiental.

Já o termo *agroquímico* tem 452 ocorrências em 250 documentos; *biocida* 49 ocorrências em 34 documentos; *plaguicida* 843 ocorrências em 225 documentos; *pesticida* 672 ocorrências em 372 documentos; *producto fitosanitario* 62 ocorrências em 45 documentos. Esses dados são sistematizados na tabela seguinte:

Tabela 5. Comparação de ocorrências dos termos nos *corpus* de estudo e no *corpus* de referência

	<i>Corpus agrotóxico</i>	<i>Corpus de referência</i>
<i>Agrotóxico</i>	-	168
<i>Agroquímico</i>	234	452
<i>Biocida</i>	692	49
<i>Plaguicida</i>	2.169	843
<i>Pesticida</i>	74	672
<i>Producto Fitosanitario</i>	999	62

Uma particularidade é que no *corpus* de referência aparece o termo *agrotóxico*. O termo *plaguicida* é o termo mais recorrente dentre os termos estudados em ambos os *corpora*. No entanto, as ocorrências de *biocida*, *pesticida* e *producto fitosanitario* se diferem: dois termos (*biocida* e *producto fitosanitario*) têm mais ocorrências no *corpus* de estudo e poucas no *corpus* de referência, e um termo (*pesticida*) tem pouca ocorrência no *corpus* de estudo e ocorrência significativa no *corpus* de referência. No entanto, é preciso sempre ter em mente que os termos analisados devem ser estudados nos textos especializados. Desse modo, suas ocorrências e seus respectivos contextos devem ser coletados nos textos legislativos, dado que buscamos equivalentes em espanhol para termos identificados em português em textos legislativos. O *corpus* de referência serviu como um contraste, inclusive, a pouca ocorrência nesse *corpus* e ocorrência significativa nos *corpora* especializados pode indicar que os termos em questão são mais especializados, uma vez que o *corpus* de referência inclui textos não especializados. Por fim, destacamos que *plaguicida* tem uma grande ocorrência em ambos os *corpora*, uma vez que é um termo mais genérico.

Após a identificação dos termos e de suas definições nos textos especializados, buscamos organizá-los em mapas conceituais discutidos na seção seguinte.

## 2.1. Mapas conceituais

Destacamos que os mapas conceituais foram construídos a partir das definições encontradas nos textos especializados. O primeiro mapa conceitual apresentado é o dos termos analisados: *plaguicida*, *biocida*, *agroquímico*. Conforme indicamos foram excluídos os termos *pesticida*, devido a sua pouca ocorrência, e *producto fitosanitario*, pelos contextos que se contradizem:

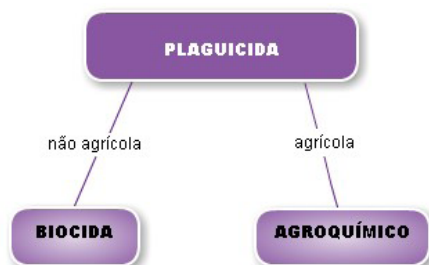


Figura 3. Mapa conceitual dos termos analisados

Fonte: as autoras



A partir dos dados obtidos nos *subcorpora* e apresentados anteriormente, *plaguicida* foi considerado o termo superordenado, enquanto que *plaguicida* de uso não agrícola seria *biocida* e, de uso agrícola, *agroquímico*. Como na legislação brasileira o contexto de *agrotóxico* é agrícola, foi desconsiderado o termo *biocida* e realizado um segundo mapa conceitual, incluindo os termos considerados equivalentes:

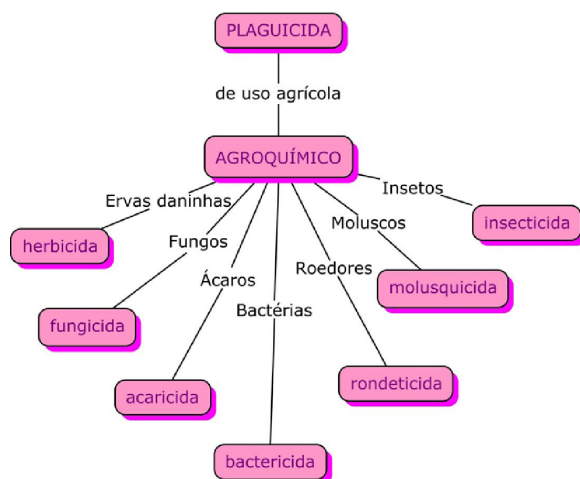


Figura 4. Mapa conceitual dos termos equivalentes

Fonte: as autoras

Além de apresentar o mapa conceitual com os equivalentes de *agrotóxico* (*plaguicida* e *agroquímico*), também estão presentes no mapa seus tipos, isto é, outros termos que apareceram nos *corpora* e se relacionam com a área do Direito Ambiental. O agroquímico que mata erva daninhas é o *herbicida*; o que mata fungos, *fungicida*; o que mata ácaros, *acaricida*; o que mata bactérias, *bactericida*; o que mata roedores, *rondeticida*; o que mata moluscos, *molusquicida*, e o que mata insetos, *insecticida*.

### 3. Considerações finais

Pelas análises realizadas, podemos concluir que as definições dos dicionários não aportaram informações suficientes para o estabelecimento dos equivalentes em língua espanhola do termo *agrotóxicos*. As prováveis

primeiras conclusões após a leitura das definições dos glossários não coincidiram com as conclusões da análise. No entanto, foi uma etapa importante, principalmente porque permitiu identificar o termo *biocida*. Apesar de serem dicionários e glossários especializados, eram de especialidades diferentes, o Meio Ambiente e a Ecologia, não condizendo exatamente com a realidade da área estudada no presente artigo, o Direito Ambiental. Por esse motivo, *pesticida*, *plaguicida* e *biocida* apareciam como sinônimos em algumas das obras consultadas, mas essa informação não pôde ser comprovada nos *corpora*; e também por esse motivo o termo *pesticida* parecia o mais utilizado nas obras, mas teve pouca ocorrência nos *corpora* dos textos legislativos.

A análise dos dados permitiu reconhecer os termos *agroquímico* e *plaguicida* como equivalentes de *agrotóxico* na área do Direito Ambiental. Ao ser flexível, a noção de equivalência funcional permite esse tipo de decisão, pois questiona a ideia de que haverá sempre um equivalente correto de uma língua a outra que será utilizado para todos os textos, em diferentes funções, áreas de saber e contextos. Apesar de o termo *agroquímico* ser mais específico para o meio agrícola, o termo *plaguicida* é muito utilizado nos mesmos contextos nos textos analisados, demonstrando serem sinônimos.

As terminologias de uma área são representativas do conhecimento especializado, um meio de expressão e comunicação profissional: os termos transmitem e representam conteúdos próprios de cada área. Não se trata de um conjunto encapsulado de informação, mas sim uma seleção específica de características semânticas segundo as condições de cada situação de uso. Deste modo, a sinonímia não prejudicaria a precisão conceitual nas comunicações profissionais, visto que seria um discurso produzido por uma comunidade discursiva específica, os especialistas se entenderiam entre si. A sinonímia não deveria ser vista como um empecilho no discurso especializado, talvez ao invés de se tentar eliminar os sinônimos, fosse interessante estudar os sinônimos existentes e em que situações comunicativas eles são utilizados. Além disso, as sinonímias podem apresentar diferentes pontos de vista, como pôde ser observado com a escolha de utilizar *agrotóxico*, *biocida*, *praguicida*, *pesticida*.

Nesse sentido, a Teoria Comunicativa da Terminologia acrescentou muito na análise visto que não se trata de uma Terminologia relacionada à padronização ou prescrição, mas sim à descrição. Em uma Terminologia descritiva, os termos são observados no uso, ou seja, nos textos especializados. Essa possibilidade de observar os termos nos contextos para analisá-los e descrevê-los é compatível com os princípios propostos pela Linguística de *Corpus*.

Por fim, construir mapas conceituais ajuda a organizar os termos de uma área do conhecimento e perceber as relações que estabelecem entre si.

Quanto às duas análises possíveis de um *corpus* citadas anteriormente, a do tipo qualitativo e quantitativo, mostra-se a necessidade de utilizar ambas e realizar uma pesquisa qualiquantitativa. Para descobrir os equivalentes de *agrotóxico*, não bastava contar frequências dos termos ou analisar somente seus contextos. As frequências e os contextos juntos facilitaram a descrição dos termos e permitiram encontrar os equivalentes buscados. Além disso, sabe-se que em Terminologia a frequência de uso de um termo não é suficiente, pois palavras com baixa frequência também têm sua importância para a área e podem ser igualmente consideradas termos e ser analisadas. Para exemplificar, lembramos que um dos termos menos frequentes dentre os termos estudados – *agroquímico* – acabou sendo eleito como equivalente de *agrotóxico* juntamente com o termo mais frequente – *plaguicida*.

É importante destacar que, apesar de a análise dos *subcorpora* apresentar números que permitem mostrar a quantidade de ocorrências e de palavras, a análise em si não depende exclusivamente das ferramentas, é uma atividade humana, podendo depender da interpretação de quem analisa os *corpora* e dos próprios textos que os compõem. Dito isso, é possível analisar os dados por diferentes perspectivas e critérios, por exemplo, analisar a relação dos termos e a equivalência de *agrotóxico* levando em consideração um país hispanofalante somente, o que incorreria em uma maneira de organizar os conceitos de forma diferente.

Os contextos de uso, o emprego dos termos, a equivalência e a relação sinônima somente puderam ser analisados levando em consideração a área do saber pretendida; em outra área, a análise poderia ser totalmente diferente. Há uma necessidade de se observar esses fenômenos em cada área específica do saber, visto que cada uma tem seu modo próprio de expressar seu fazer científico. Esse tipo de análise, bem como o contraste com a definição de dicionários gerais, pode ser realizado em estudos futuros, o que trará novos olhares e lançará luzes sobre os resultados aqui apresentados.

## Referências

- Allaby, M. (1984). *Diccionario del medio ambiente*. Madrid: Pirâmide.
- Aluísio, S. M. & Almeida, G. M. B. (2006). O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio*, 4(3), 156–178.

- Araújo, M. (2010). Terminologia e sinonímia: são os sinônimos indesejáveis nos discursos especializados? In N. A. Isquerdo & M. J. B. Finatto (Eds.), *As ciências do léxico: Lexicologia, Lexicografia e Terminologia v. IV*, (pp. 519–537). Campo Grande: Ed. UFGS; Editora da UFRGS.
- Berber Sardinha, T. (2004). *Linguística de Corpus*. Barueri: Manole.
- Berber Sardinha, T. (2000). Linguística de Corpus: Histórico e problemática. *DELTA*, 16(2), 323–367.
- Bevilacqua, C. R. (2013). Por que e para que a Linguística de Corpus na Terminologia. In S. Tagnin & C. R. Bevilacqua (Eds.), *Corpora na Terminologia* (pp.11–27). São Paulo: HUB Editorial.
- Bevilacqua, C. R., Maciel, A. M. B., Reuillard, P. C. R., Scheren, C. & Kilian, C. K. (2013). Combinatórias léxicas da linguagem legislativa. In C. Murakawa & O. L. Nadin (Eds.), *Terminologia: uma ciência interdisciplinar* (pp. 227–244). São Paulo: Cultura Acadêmica.
- Cabré, M. T. (2005). La Terminología, una disciplina en evolución: Pasado, presente y algunos elementos de futuro. *Debate Terminológico*, n.1.
- Cabré, M. T. (2004). A Terminologia hoje: Concepções, tendências e aplicações (S. Kerschner, Trad.). *Cadernos de Tradução*, n. 17.
- Cabré, M. T. (2002). Terminología y Lingüística: La teoría de las puertas. *Estudios de Lingüística del Español*, v. 16.
- Cabré, M. T. (1999). *La terminología: representación y comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.
- Cabré, M. T. (1993). *La terminología: teoría, metodología, aplicaciones*. Barcelona: Empuries.
- Costa, M. I. P. (2009). *Estudo Preliminar da terminologia empregada pela polícia civil do RS no Boletim de Ocorrência policial*. Dissertação de mestrado, Universidade Federal do Rio Grande do Sul, Porto Alegre.
- Gémar, J. C. (1998). Les enjeux de la traduction juridique. Principes et nuances. Disponível em: <<http://www.tradulex.com/Bern1998/Gemar.pdf>>. Consultado em: 20 abril 2017.
- Larousse (1988). *Diccionario Ilustrado de las Ciencias*. Indiana: Larousse.
- Lei nº 7.802, de 11 de julho de 1989. Dispõe sobre a pesquisa, a experimentação, a produção, a embalagem e rotulagem, o transporte, o armazenamento, a comercialização, a propaganda comercial, a utilização, a importação, a exportação, o destino final dos resíduos e embalagens, o registro, a classificação, o controle, a inspeção e a fiscalização de agrotóxicos, seus componentes e afins, e dá outras providências. Disponível em: [http://www.planalto.gov.br/ccivil\\_03/leis/L7802.htm](http://www.planalto.gov.br/ccivil_03/leis/L7802.htm).
- Maciel, A. M. B. (2013) Terminologia e Corpus. In S. Tagnin & C. R. Bevilacqua (Eds.), *Corpora na Terminologia* (pp. 29–45). São Paulo: HUB Editorial.
- Mariscotti, E. T. P. (1993). *Glosario sobre Ecología y Medio Ambiente*. Buenos Aires, Argentina: Orientación Grafica Editorial.

- Martín, A. & Santamaría, J. M. (2000). *Diccionario terminológico de contaminación ambiental*. Navarra: EUNSA.
- Medicci, E. G. & Vivas, J. O. O. (1979). *Glosario Ambiental*. Caracas, Venezuela: Ediciones del Congreso de la República.
- Moragas, W. M. & Scheneider, M. O. (2003). Biocidas: suas propriedades e seu histórico no Brasil. *Caminhos de Geografia*, 3(10), 26–40.
- Parra, F. (1984). *Diccionario de ecología, ecologismo y medio ambiente*. Madrid: Alianza Editorial.
- Pérez Hernández, M. C. (2002). Explotación de los corpórea textuales informatizados para la creación de bases de datos terminológicas. *Estudios de Lingüística del Español*, vol. 18.
- Rodríguez, E. P. (1992). *Diccionario Ecológico Ilustrado*. Bogotá: Espacio Editorial.
- Wüster, E. (1998). *Introducción a la teoría general de la terminología y a la lexicografía terminológica*. Trad. de A. C. Nokerman. Barcelona: Universitat Pompeu Fabra.

[recebido em 29 de março de 2018 e aceite para publicação em 15 de novembro de 2018]

# **QUANDO O LÉXICO DÁ BANDEIRA – ASPECTOS COGNITIVO-DISCURSIVOS DA MUDANÇA SEMÂNTICA NA CONSTRUÇÃO DE BRASILEIRISMOS EM REGISTROS LEXICOGRÁFICOS LUSO-BRASILEIROS**

WHEN THE LEXICON SHOWS THE COLOR – DISCURSIVE-  
COGNITIVE ASPECTS OF SEMANTIC CHANGE IN THE  
CONSTRUCTION OF BRAZILIANISMS IN LUSO-BRAZILIAN  
LEXICOGRAPHIC REGISTERS

Anderson Salvaterra Magalhães\*  
asmagalhaes@unifesp.br

Janderson Lemos de Souza\*  
janderson.souza@unifesp.br

Neste artigo, busca-se demonstrar a adequação de articular princípios da Análise Dialógica do Discurso – um campo do conhecimento que emerge da recepção brasileira ao pensamento do Círculo Bakhtin-Medvedev-Voloshinov (Círculo BMV) – com fundamentos da Linguística Cognitiva para o tratamento de questões morfossemânticas do léxico do português brasileiro que indicam importantes atos na política lusófona. Especificamente, perseguem-se dois objetivos: 1) identificar condições cognitivo-discursivas próprias do português brasileiro que impactam seu estatuto vernáculo e 2) descrever um caso de mudança semântica que ilustra novas conceptualizações no léxico interno e registros do léxico externo a serviço de um projeto lexicográfico caracterizado como ato responsável (Bakhtin). Para isso, selecionam-se dos primeiros trabalhos lexicográficos luso-brasileiros, que datam dos séculos XVIII e XIX, duas unidades simbólicas em que constam tensões conceituais entre o lusitano e o brasileiro, a saber, *bandeira* e *bandeirante*. Os primeiros registros são cotejados com representativo trabalho lexicográfico brasileiro e lusitano do século XXI para fins de identificação, descrição e análise da mudança semântica que produz o senso de ‘brasileirismo’ tanto no Brasil quanto em Portugal a partir de *frames* (Fillmore) lusitanos. A discussão dos dados indica que a dimensão vernacular brasileira se constrói não por uma perspectiva propriamente brasileira, mas lusitana face às conceptualizações das relações sociais travadas em terras americanas no período colonial.

---

\* Universidade Federal de São Paulo, Brasil.

**Palavras-chave:** Mudança semântica. Léxico brasileiro. Lexicografia e política lusófona. Dialogismo. Semântica cognitiva.

In this article, the general aim is to demonstrate the appropriateness of articulating principles of Dialogic Discourse Analysis – a field of knowledge which derives from the Brazilian reception to the Bakhtin-Medvedev-Vološinov Circle's thought (BMV Circle) – with fundamentals of Cognitive Linguistics to tackle morphosemantic issues of the Brazilian lexicon which indicate important acts in the Lusophone politics. Two specific aims are pursued: 1) to identify the cognitive-discursive conditions typical of the Brazilian Portuguese which influence its vernacular status; 2) to describe a case of semantic change which illustrates new conceptualizations in the internal lexicon and registers of the external lexicon at the service of a lexicographic project which is characterized as a responsible act (Bakhtin). In order to reach those goals, two symbolic units – *bandeira* and *bandeirante* – were selected from the first Luso-Brazilian lexicographic works, which date from the 18<sup>th</sup> and 19<sup>th</sup> centuries. Those first registers are compared with representative Brazilian and Portuguese lexicographic works of this century in order to identify, describe and analyse the semantic change which produces the sense of 'Brazilianism' from Lusitanian frames both in Brazil and in Portugal. The discussion of the data indicates that the Brazilian vernacular dimension is not built from an actually Brazilian perspective, but from a Lusitanian one, considering their conceptualizations of the social relations established in America in the colonial period.

**Keywords:** Semantic change. Brazilian lexicon. Lusophone lexicography and politics. Dialogism. Cognitive semantics.



## 1. Introdução

Neste artigo, busca-se demonstrar a adequação de articular princípios da Análise Dialógica do Discurso – um campo do conhecimento que emerge da recepção brasileira ao pensamento do chamado Círculo de Bakhtin (Brait & Magalhães 2014) – com fundamentos da Linguística Cognitiva (Fillmore 1982; 1985; Langacker 1987; 1991; 1994; 2008; 2009) para o tratamento de questões morfossemânticas do léxico do português brasileiro.

O problema aqui tratado escapa a amarras estritamente formais e se coloca numa arena de tensão conceitual, dado que o limiar entre o que é propriamente brasileiro em contraponto ao que é português esbarra em

questões linguísticas, sim, mas igualmente sociais, políticas e culturais. Assim, dois aspectos são assumidos como premissas na presente discussão: 1) a política lusófona no Brasil gerou uma ferida histórico-cultural constitutiva da identidade linguística nacional tensionada pela língua no/do Brasil (Dias 1996; Fiorin 2009; Magalhães 2012; 2013; Schneiders 2017) e 2) a produção de vocabulários e dicionários tem funcionado como tecnologia de uma determinada memória do português no Brasil (Auroux 1992; Barros 2000; Magalhães 2015; Petri & Medeiros 2013).

A primeira premissa sustenta a ideia de que a língua que faz sentido da brasilidade é, em tese, de outrem; é lusitana. Demais línguas com as quais o português entrou em contacto em terras brasileiras funcionaram como aportes, mas não tiveram força política para emplacar como idioma do Brasil. Afinal, nem as diferentes etnias indígenas, nem as etnias africanas trazidas escravizadas para o Brasil (*cf.* Peter 2015) visualizavam nesse espaço geográfico uma unidade política, como o colonizador. Parece consequente que a língua de quem projetou tal unidade acabe por se definir como língua dessa projeção.

Um dos problemas que daí resultam é o modo como o valor lusitano habita o falar brasileiro. Em certos casos, como o que é analisado neste artigo, o lugar histórico-social lusitano de fazer sentido constitui a memória da língua na condição de ‘brasileirismo’ tanto no falar brasileiro como no europeu (Magalhães 2012; 2013). Em outras palavras, a língua tomada como vernáculo brasileiro guarda um ponto de vista de outrem a partir do qual se produz sentido, e o léxico – relevante instância da língua em que se manifestam conceptualizações características das relações sociais travadas na comunidade linguística – dá pistas de especificidades cognitivo-discursivas do português produzido como idioma brasileiro. Mas cabe a pergunta: como pode um juízo lusitano caracterizar o vernáculo brasileiro? O caso aqui recortado indica possível resposta à indagação.

A segunda premissa trata dos mecanismos materiais que favoreceram e favorecem a construção e a consolidação do português como vernáculo brasileiro e não mera transposição de um idioma europeu para terras americanas. A produção de vocabulários e de dicionários, ainda que no afã de defender o idioma como “língua brasileira”, dá pistas acerca desse lugar de alteridade de onde se faz sentido no português do Brasil. Tais pistas são deixadas tanto na macroestrutura quanto na microestrutura dos produtos lexicográficos, isto é, tanto naquilo que é selecionado para constar no vocabulário ou dicionário quanto naquilo que é selecionado para ser apresentado de cada item e como é apresentado (Biderman 2003; Geeraerts 2006a; 2006b).



Se nos projetos lexicográficos não são encontrados documentos fidedignos do português efetivamente em uso desde o século XVI no Brasil (Biderman 2003), o cruzamento da macro e da microestrutura dos vocabulários e dicionários de diferentes séculos auxilia no desenho do(s) propósito(s) pragmático(s) dos projetos editoriais (Geeraerts 2006a) a delinear, por exemplo, os implícitos em conformidade com os destinatários pretendidos. Assim, procede-se ao esboço dos processos cognitivo-discursivos pinçados que acabam por fomentar a política linguística brasileira. Portanto, o registro lexicográfico não é tomado como documentação do uso, mas como documentação de discursos sobre o uso que ordenam o modo como aqueles que implementaram e implementam uma política linguística processam as relações sociais, linguísticas e culturais estabelecidas no Brasil.

Com base nessas duas premissas, perseguem-se dois objetivos específicos inter-relacionados: 1) identificar condições cognitivo-discursivas próprias do português brasileiro que impactam seu estatuto vernáculo e 2) descrever um caso de mudança semântica que atua em novas conceptualizações no léxico interno e transparece em registros do léxico externo a serviço de um projeto lexicográfico caracterizado como ato responsável (Bakhtin 2010c).

Para alcançar estes objetivos, com base na orientação metodológica de Bakhtin (2010a) de integrar os estudos linguísticos aos estudos dialógicos (neste trabalho, reconhecidos pela produção em *Análise Dialógica do Discurso*) e no entendimento de Langacker (2008) de que as funções que moldam e restringem a língua incluem a função semiológica de simbolizar conceptualizações por meio de sons e gestos (a caracterizar a gramática cognitiva como semiológica e multimodal), destacam-se as relações linguístico-discursivas entre uma unidade simbólica – *bandeira* – e outra a partir dela formada – *bandeirante* – em que constam as tensões conceituais entre o lusitano e o brasileiro. Trata-se de unidades simbólicas cuja semiose remonta um episódio da historicização do Brasil – historicização compreendida como processo de fazer sentido da história e que se dá não só por meio, mas por causa da língua que ajusta um grupo social.

A construção e discussão do caso mobiliza um dispositivo analítico que recupera principalmente o conceito de *memória da língua* (Bakhtin 2003; 2010b) e os articula com fundamentos da Linguística Cognitiva, especialmente a semântica de *frames*, empreendida por Fillmore (1982; 1985), quando se deixou de distinguir *frame* de *cena* (cf. Fillmore 1977b), e a gramática cognitiva, tal como proposta por Langacker (1987; 1991; 1994; 2008; 2009), por conta da noção de construção ali sustentada, cujas bases semiológicas e multimodais foram assinaladas acima.

Trata-se de abordagens centradas no significado, que, entretanto, lançam luz sobre diferentes aspectos do fenômeno linguístico. O dispositivo da Análise Dialógica do Discurso dá acesso às questões semiótico-ideológicas; a Semântica de *Frames* instrui a concepção de léxico adotada, conforme detalhado adiante; e a gramática cognitiva de Langacker (1987; 1991; 1994; 2008; 2009), por conceber que construções gramaticais gozam de diferentes graus de rotinização e convencionalidade numa comunidade linguística, permite tratar *bandeira* e *bandeirante* como unidades simbólicas, seja como itens do léxico interno (pertinente à continuidade entre a gramática e o léxico), seja como itens do léxico externo (pertinente à lexicografia).

Dado que todas as vertentes da Linguística Cognitiva são lexicalistas porque são construcionais (cf. Traugott & Trousdale 2013), exige-se tanto de descrições internas à Linguística Cognitiva quanto, e mais ainda, de discussões teóricas que, como neste artigo, se disponham a aproximar diferentes quadros teóricos em que um deles seja a Linguística Cognitiva que especifiquem a concepção de léxico e o modelo construcional adotado. O modelo aqui adotado é a Gramática Cognitiva, em que “(...) o léxico e a gramática formam uma gradação que consiste tão-somente em grupos de estruturas simbólicas” (Langacker 2008, p. 5 – tradução nossa).

A condição de itens do léxico externo (verbetes) serve como pista para a identificação da conceptualização que se pretende capturar na condição de unidades simbólicas, ressalvado o salto que há entre uma condição e outra. Nos verbetes, vislumbra-se o universo semântico-cultural que, referendado pela política linguística promovedora de instrumentos específicos, funciona como vetor de discursos sobre a língua, suas normas e usos (Barros 2000) a incidir sobre as conceptualizações.

Foram consultados os seguintes vocabulários e dicionários luso-brasileiros: a) *Diccionario da Lingua Portuguesa*, de D. Rafael Bluteau, revisto por Antonio de Moraes Silva, 1789; b) *Diccionario da Lingua Brasileira*, de Luiz Maria da Silva Pinto, 1832; c) *Vocabulário Brasileiro – para servir de complemento aos dictionarios de lingua portuguesa*, de Braz da Costa Rubim, 1853; d) *Grande Dicionário Houaiss da Língua Portuguesa* [em linha], 2012; e) *Dicionário infopédia da Língua Portuguesa com Acordo Ortográfico* [em linha], 2003-2017. Evidentemente, não se trata de um levantamento exaustivo. Essas obras configuram os primeiros e uns dos mais recentes registros lexicográficos dos vocábulos estudados.

A discussão segue com outras três principais secções além desta introdução. Na primeira, discutem-se as condições cognitivo-discursivas para o desenho de uma produção linguística que possa ser considerada

vernacular. Aí é construído o caso em torno dos brasileirismos, categoria pertinente ao léxico externo. Na segunda, discute-se o caso de *bandeira* e *bandeirante*, pontuando aspectos históricos e cognitivos que fomentam a mudança semântica que leva à tensão vernacular flagrada, pertinentes ao léxico interno. Por fim, à guisa de conclusão, reapresentam-se as cicatrizes simbólicas flagrantes no caso destacado para problematizar as condições da própria conceituação do que seja vernáculo e brasileiro.

## 2. Condições cognitivo-discursivas de brasilidade e o vernáculo do Brasil

Desde a chegada dos portugueses a terras posteriormente reconhecidas e organizadas como Brasil, o estatuto do português se alterou significativamente. *Grosso modo*, identifica-se como língua do colonizador, língua da colônia, língua do império e, já na República, língua nacional. Essa trajetória muito dista do percurso do português na Europa.

No período colonial, as relações entre diferentes etnias implicaram o contacto de diferentes visões de mundo, e as línguas envolvidas nessas relações dão pistas acerca dos modos de conceber o que se processava naquela(s) sociedade(s). O conceito e projeto de *colônia* era lusitano, e não indígena, sendo óbvio que o europeu se definia como colonizador e definia o indígena como colonizado. Mas o que dizer do ponto de vista indígena sobre essas relações? Esse ponto de vista deixou pistas linguísticas? O modo de contar a história do Brasil parece mobilizar uma memória que difere da do autóctone. Ademais, a condição de escravizada das etnias africanas trazidas para essa colônia também promovia modos díspares de conceber o que se passava.

Com a política linguística de D. João VI no século XVIII a favor do português em detrimento das línguas gerais e com o estatuto sempre subalterno legado às línguas africanas, o português logrou condição de língua oficial e de uso no Brasil. As línguas indígenas e africanas ficaram restritas a comunidades destacadas da cultura sociopolítica que se instituíam. Entretanto, a língua portuguesa que se estabeleceu na Colônia e se confirmou no Império e na República guardou índices do que se passou especificamente nesse espaço.

Esse modo de compreender a relação entre língua e sociedade alinha-se com a perspectiva dialógica de linguagem, segundo a qual “a linguagem não é um dom divino nem um presente da natureza. É o produto da atividade

humana coletiva e reflete em todos os seus elementos tanto a organização econômica como a sociopolítica da sociedade que a gerou” (Voloshinov 2013, p. 141 – grifos da edição consultada). Nessa abordagem, a língua não figura como entidade destacada do grupo social, mas é ao mesmo tempo fruto desse grupo e condição desse grupo. Essa compreensão teórica se coaduna com a ideia de *sociação* de Simmel (2006, p. 61), para quem a base de qualquer sociedade humana é:

a forma (que se realiza de inúmeras maneiras distintas) na qual os indivíduos, em razão de seus interesses – sensoriais, ideais, momentâneos, duradouros, conscientes, inconscientes, movidos pela causalidade ou teleologicamente determinados – se desenvolvem conjuntamente em direção a uma unidade no seio da qual esses interesses se realizam.

Por esta leitura, é possível afirmar que língua e sociação condicionam-se mutuamente e que o vernáculo se define pelos processos linguísticos gerados nas relações sociais de determinado grupo e pelos produtos verbais dessas relações. É possível afirmar também que a língua não se destaca da ideologia. Por ser este um conceito bastante disputado, vale recuperá-lo do pensamento dialógico.

Segundo Voloshinov (2013, p. 138), ideologia é “o conjunto de reflexos e interpretações da realidade social e natural que *se sucedem no cérebro do homem*, fixados por meio de palavras, desenhos, esquemas ou outras formas *sígnicas*” (grifos da edição consultada). Assim, a despeito do que haja de natural na linguagem, dialogicamente, a língua não pode ser reduzida a um processamento imediato de relações físico-naturais. Ao conceituar o universo físico ao seu redor, o indivíduo produz sentido e instala-se na cultura, alçando-se à condição de sujeito num grupo social.

Já na Linguística Cognitiva, o conceito de ideologia não goza de *status* destacado dentre os fatores que condicionam os processos cognitivos. Pode-se apenas inferir da condição da cognição como *situada* (Croft & Cruse 2004) e como *distribuída* (Langacker 1994; Silva 2009) que os fatores social, histórico e cultural incluem o ideológico.

Especialmente no que tange ao fator social, sua influência se faz sentir desde propostas iniciais de nomear a teoria como linguística sociocognitiva até a recente “(...) institucionalização da nova área da sociolinguística cognitiva, como extensão e linha de investigação em linguística cognitiva (...)”, que inclui em seu campo a “(...) investigação sobre modelos cognitivos culturais subjacentes a atitudes linguísticas e políticas de língua e investigação

sobre ideologias sócio-políticas e socioeconômicas” (Silva 2009, p. 192). Dessa forma, uma das características definidoras do *frame* para Fillmore (1977a – *cf. infra*), a perspectiva, recebida por Langacker (1987; 1991; 1994; 2008; 2009) para formular a conceptualização como dinâmica e perspectivizada, passa a ser invocada para explicar a variação sociolinguística, considerada como:

(...) forma específica de significado, mais precisamente, diferentes tipos de significado não denotacional: significado emotivo (de termos pejorativos, por exemplo), significado social (de termos regionais e sociais), significado estilístico (de termos populares e eruditos) e significado pragmático-discursivo (único de expressões como as interjeições e os marcadores discursivos; presente em termos como senhor, você, tu e outras formas de tratamento). (Silva 2009, pp. 193–194)

Langacker (1994) recupera o caráter distribuído da cognição para estabelecer a versão de relativismo que instrui a gramática cognitiva, a de que a cognição inclui a língua e a cultura: “(...) a existência dos três termos distintos *língua*, *cognição* e *cultura* não deveria nos levar ao engano de pensar que essas são entidades separadas, não sobrepostas” (p. 26 – tradução nossa).

No presente artigo, o caso que se descreve se constrói justamente na relação língua-cultura. A atenção tanto ao caráter situado quanto ao caráter distribuído da cognição, característica da Linguística Cognitiva, associada à concepção de cognição como constituída pela língua e pela cultura, característica da Gramática Cognitiva, leva a afirmar que, ao conceituar em língua portuguesa algo próprio das relações travadas no Brasil, ainda que o valor social seja lusitano, o processamento linguístico e o produto verbal registram fragmentos da história realizada nesse espaço geopolítico e, assim, caracterizam a língua ‘(d)aqui’ por um *frame* a partir de uma perspectiva ‘de lá’. E o registro do léxico externo, desse ponto de vista teórico, figura como rastro dessa relação entre língua e sociedade porque guarda indícios da história do grupo social que o mobiliza.

Dialogicamente, o registro do léxico externo institui um *ato responsável*, socioculturalmente circunscrito (Bakhtin 2010c; Sobral 2006) no âmbito da política linguística brasileira. É responsável por, simultaneamente, responder às circunstâncias socioculturais em que tais instrumentos se circunscrevem e por implicar a responsabilidade político-linguística de tal resposta. Nesses termos, a condição de ato não mistura uso e menção, mas confere diferente estatuto a um e a outro. O uso corrente é um ato de comunicação que atende a determinado objetivo interacional. A menção constitui um ato metalinguístico que atende a determinado objetivo cultural.

A distinção tradicional entre *uso* e *menção* assume caráter metodológico na Linguística Cognitiva, sem implicar distinção quanto aos processos cognitivos que organizam ambas as formas de uso da língua, o que se permite considerar como outra compatibilidade entre os dois quadros teóricos aqui aproximados: a Análise Dialógica do Discurso e a Linguística Cognitiva.

Para tratar da questão levantada acerca de brasileirismos, aciona-se a Semântica de *Frames*. Foi assinalado que o conceito de *frame* muda na obra de Fillmore. Os fatores *orientação* e *perspectiva* são considerados por Fillmore (1977a) como definidores do *frame*. Os conceitos de *frame* e *cena* são considerados distintos por Fillmore (1977b). O conceito de *frame* passa a abarcar o de *cena* a partir de Fillmore (1982), que, ao “(...) enfatizar as continuidades, e não as descontinuidades, entre a linguagem e a experiência” (p. 111 – tradução nossa), estabelece a definição de *frame* como “(...) qualquer sistema de conceitos relacionados de tal forma que, para entender qualquer um deles, é preciso entender toda a estrutura na qual ele se encaixa (...)” (p. 111 – tradução nossa). É essa a definição de *frame* que inaugura um projeto lexicográfico, o *FrameNet*, do qual não trataremos aqui (cf. Fillmore & Atkins 1992).

Como o próprio autor explicitaria mais tarde, a Semântica de *Frames* trata “da relação entre textos linguísticos, o contexto em que são instanciados e o processo e os produtos de sua interpretação” (Fillmore 1985, p. 222 – tradução nossa), o que o motiva a classificá-la como uma Semântica da Compreensão.

A partir daí, recupera-se Langacker (2008), que também parte do conceito de *frame* fornecido por Fillmore (1982); reconhece a influência desse conceito em outra abordagem cognitivista, caracterizada por Traugott e Trousdale (2013) como gramática das construções cognitiva, na qual o *frame* é um dos componentes de um modelo cognitivo idealizado; e conclui por empregar o conceito de domínio: “*Domínio* tem maior generalidade, uma vez que nem *frame* nem MCI se aplicam muito bem a domínios básicos (ex.: tempo ou espectro de cores). Um *frame* pode ser grosseiramente comparado a um domínio não básico” (Langacker 2008, pp. 46–47).<sup>1</sup>

---

1 Texto no original: “*Domain* has the greatest generality, since neither *frame* nor *ICM* applies very well to basic domains (e.g. time or color space). A *frame* may be roughly comparable to a nonbasic domain.” (tradução nossa)

Incidentalmente, cabe registrar que a principal razão por que Dancygier e Sweetser (2014) preferem descrever a metaforicidade por meio de *frames*, e não de domínios, é exatamente a generalidade do conceito de domínio por entenderem que a metáfora se dá num nível mais específico de correspondência (*mapping*), o dos *frames*. Portanto, há consenso quanto ao teor de cada conceito, seja para aderir, seja para rejeitar.

Na análise pretendida neste artigo, mobilizam-se o conceito de *frame* pela especificidade que permite discernir entre o brasileiro e o lusitano sob a mesma língua, não segundo a mesma conceptualização, e a definição de construção gramatical segundo Langacker (2008; 2009):

Uma construção é definida como uma expressão (de qualquer tamanho) ou como um esquema abstraído de expressões para capturar o que lhes é comum (em qualquer nível de especificidade). As expressões e os padrões que instanciam são, portanto, iguais em sua natureza básica, a diferir apenas no grau de especificidade. Tanto expressões específicas quanto esquemas abstraídos são capazes de serem rotinizados psicologicamente e convencionalizados em uma comunidade de fala, caso no qual constituem unidades linguísticas estabelecidas. (Langacker 2009, p. 2)<sup>2</sup>

Portanto, a relação entre a gramática e o léxico corresponde à relação entre o esquema e a expressão. Assim, *bandeira* e *bandeirante* assumem o *status* de expressões (instanciações de construções) no léxico interno e o *status* de verbetes na tensão acerca do vernáculo enfocada neste artigo.

Magalhães (2015), ao discutir a formação do vocábulo *botocudo* e suas pistas vernaculares no uso e na tarefa lexicográfica, demonstra como uso e menção participam da política lusófona no Brasil. O autor demonstra que a dimensão ideológica implicada no semiótico é o que gera diferentes *memórias* (Bakhtin 2003, p. 380) da língua portuguesa no Brasil e alhures, o que distingue entre o português tomado como vernáculo brasileiro e o português de outrem. Isso significa que as formas linguísticas não guardam uma relação direta com o meio, mas guardam uma memória do que se rotiniza em atos responsáveis (Bakhtin 2010a; 2010b; Magalhães 2015) e, a partir daí, do que registrar no léxico externo.

Neste artigo, *bandeira* e *bandeirante* são tomados como unidades simbólicas com diferentes graus de analisabilidade (propriedade gradiente relativa à identificação das formas simples que participam da instanciação de uma forma mais complexa) e composicionalidade (propriedade gradiente relativa à contribuição semântica das formas simples para o significado de uma forma mais complexa, além da contribuição semântica da própria construção, do esquema), assim como verbetes cujas definições evidenciam diferentes

---

2 Texto no original: "A construction is defined as either an expression (of any size), or else a schema abstracted from expressions to capture their commonality (at any level of specificity). Expressions and the patterns they instantiate are thus the same in their basic nature, differing only in degree of specificity. Both specific expressions and abstracted schemas are capable of being entrenched psychologically and conventionalized in a speech community, in which case they constitute established linguistic units." (tradução nossa)



conceptualizações ao longo do tempo e o convívio com outras expressões, o que exige tratar aqui desses dois fenômenos. As conceptualizações capturadas lexicograficamente fornecem pistas sobre a mudança semântica que se pretende descrever, enquanto o convívio com outras expressões remete a outras construções gramaticais que vão configurando o léxico do português brasileiro. Portanto, são exigências diretamente relacionadas aos objetivos deste artigo.

Os fundamentos mobilizados da análise dialógica do discurso e da linguística cognitiva permitem considerar os verbetes como reflexos da política linguística brasileira, pelas acepções registradas. Não são apenas registros de usos, são sobretudo registros de processos linguísticos projetados para fins de uma política lusófona, no que constituem atos responsáveis, ou seja, atos que respondem a uma conjuntura histórico-social. Por essa razão, na tarefa lexicográfica, as anotações de regionalismos indicam valores sociais que deixam ver as fronteiras de grupos sociais, e, para a presente discussão, os brasileirismos são o foco de atenção. Essas anotações são importantes atos da política linguística que, ao ativarem determinados *frames*, fomentam determinada memória da língua.

Como defende Simmel (2006), a herança cultural é condição para que um indivíduo integre um grupo social, e a língua bem como a política linguística de uma sociedade são peças-chave nesse processo.

Mas não está em questão somente a hereditariedade em sentido puramente biológico.

*Também os elementos espirituais que se objetivaram em palavras e conhecimentos, em inclinações afetivas e normas de vontade e juízo, e que penetraram o indivíduo como tradições conscientes e inconscientes, fazem isso de maneira tanto mais segura e universal quanto mais consolidada e evidente elas tenham crescido dentro do espírito de uma sociedade que se desenvolveu ao longo do tempo – isto é, quanto mais antigas forem as tradições.* (Simmel 2006, p. 43 – grifos nossos)

A memória da língua portuguesa não apenas gravada, mas promovida pelos dicionários, constitui vetor desse legado, uma vez que cristaliza relações semânticas e dá a elas *status* de produto verbal estável. Mas nessa camuflagem ficam as pistas dos processos que contam em parte como se definiu a dimensão vernacular do português brasileiro.

Os dicionários consultados para a discussão deste artigo apresentam os itens lexicais semasiologicamente, em ordem alfabética. Isso impacta como cada item é apresentado, a influenciar os implícitos e subentendidos, ou, nos termos do dispositivo teórico aqui mobilizado, a saliência semântica.



De acordo com Geeraerts (2006c, p. 75), a saliência consiste da “reflexão estrutural de fenômenos pragmáticos”<sup>3</sup>, em outras palavras, “é o lugar onde estrutura e uso se encontram”.<sup>4</sup> O autor refina a definição esclarecendo que essa articulação do semântico e do pragmático requer a compreensão de *estrutura linguística* não apenas como um conjunto de possibilidades, mas como um conjunto de *probabilidades*. Desse ponto de vista, a macroestrutura dos projetos lexicográficos, em sua condição de reflexos de determinada política linguística, influencia a probabilidade de uso por construir uma memória sobre determinada base conceitual. Não reflete as frequências dos usos que antecedem aos registros por embaraços da própria tarefa lexicográfica mas influi sobre usos posteriores.

Geeraerts (2006c) afirma que há quatro principais tipos de saliência semântica, dos quais interessam para esta discussão três: de perspectiva (que o autor também chama de *destaque*), semasiológica e onomasiológica.

A saliência de perspectiva diz respeito às “diferenças de destaque a diferentes partes do recorte [*chunks*] da realidade extralinguística evocada por um conceito particular” (Geeraerts 2006c, p. 90).<sup>5</sup> Cognitivamente, corresponde a um elemento da mesma base conceitual subfocalizado, como o conceito de *mão* subentende o de *braço*, e destaca uma parte específica desse “recorte da realidade extralinguística”. Trata-se de uma relação de figura e fundo, relevante também na gramática cognitiva.

A saliência semasiológica diz respeito à “relação entre as várias possibilidades semânticas de um dado item lexical” (Geeraerts 2006c, p. 79)<sup>6</sup> e, sempre segundo o autor, coincide com a prototipicidade, seja prototipicidade propriamente semântica, seja prototipicidade referencial. Já a saliência onomasiológica trata das várias realizações verbais de um dado conceito.

Neste artigo, compreende-se que o vernáculo é uma dimensão político-linguística e que a produção de sentido é um processo que abarca mais do que trabalho linguístico. Isto porque fazer sentido envolve as condições de sociação e de legado cultural que, no processamento cognitivo, são descritos, entre outros, pelos *frames* ativados. Estes, por sua vez, fundamentam, por exemplo, os implícitos operantes na macro e microestruturas dos projetos lexicográficos. Assim, o registro de brasileirismos configura pista do que emerge das relações próprias do Brasil. Resta ver a partir de qual base conceitual e por meio de quais processos cognitivos.

3 Texto no original: “the structural reflexion of pragmatic phenomena.” (tradução nossa)

4 Texto no original: “the place where structure and use meet.” (tradução nossa)

5 Texto no original: “The differences of perspectival attention attached to different parts of the overall chunk of extralinguistic reality evoked by a particular concept.” (tradução nossa)

6 Texto no original: “a relationship among the various semantic possibilities of a given lexical item.” (tradução nossa)

### 3. Quando o léxico dá bandeira

Com esta base epistêmica, abordam-se os projetos editoriais de vocabulários e dicionários como instrumentos linguísticos (Auroux 1992; Barros 2000) nos termos aqui discutidos, instrumentos da política lusófona no Brasil, por meio dos quais se guardam índices daquilo que é referendado como *do* Brasil. A tarefa lexicográfica constitui, portanto, um ato responsável que se define por um conjunto de atividades, das quais se destacam: (a) seleção do que constar no dicionário, (b) decisão de como apresentar o repertório selecionado, (c) seleção do que apresentar de cada item do repertório e (d) decisão sobre como apresentar as acepções em cada item. De acordo com Geeraerts (2006a), o propósito pragmático impacta diretamente sobre tais atividades. Em termos dialógicos, o propósito pragmático cumpre discursivamente uma função na política linguística.

Essas atividades mobilizam categorias pragmático-discursivas que elucidam a quem se destinam tais obras, qual o referencial conceitual em cada uma delas e que sentidos fazem os implícitos nos atos discursivos.

Para proceder a esta discussão, destacam-se duas unidades simbólicas cuja trajetória semântica indicia sentidos limítrofes entre a dimensão ádvena e vernácula: *bandeira* e *bandeirante*. A observação dos primeiros registros luso-brasileiros e de outros bem recentes permite o cotejo de como se deu o processo de vernacularização dos verbetes selecionados.

Macroestruturalmente, a unidade simbólica mais simples (*bandeira*) consta em todos os projetos lexicográficos consultados; a unidade simbólica mais complexa (*bandeirante*) passa a constar a partir da segunda metade do século XIX. A princípio, aí está uma indicação de estabilidade semasiológica. Microestruturalmente, as acepções indicam minúcias do jogo semasiológico. O cotejo do que se apresenta em cada projeto lexicográfico deslinda os *frames* acionados para documentar as condições de produção de sentido nesses séculos. Isso dá pistas acerca do lugar social a partir do qual se conceituavam as relações culturais.

No século XVIII, período colonial, no primeiro registro lexicográfico luso-brasileiro de *bandeira*, há seis acepções além de dois registros idiomáticos (Quadro 1). Destas acepções, sob o *frame* militar, a conceptualização evolui do mais experiencial (insígnia) ao mais metonímico (companhia), o que Langacker (2009) considera como deslocamento da zona de ativação. A maneira de apresentar tais acepções não dá muitas dicas acerca de possíveis relações cognitivas internas ao *frame*. Já o *frame* agricultura é introduzido pelo sintagma preposicional “do milho”. Desconsiderando a metáfora,

é possível afirmar que o *frame* militar organiza o universo lusitano como referencial de significação. O ponto de vista para conceituar é europeu.

**Quadro 1. Bandeira: Lexicografia luso-brasileira do séc. XVIII**

Elementos macroestruturais	Elementos microestruturais	
Obra	Verbetes	Acepções
<i>Diccionario da Lingua Portuguesa</i> (1789)	Bandeira, s. f.	insignia militar, he huma peça de lenço, ou seda, com pinturas, armas, talvez quarreada de varias cores, para se conhecerem, e ajuntarem a ella os soldados, que vão debaixo dessa bandeira, ou pertencem á companhia do Chêfe, cuja he a bandeira; nos navios tambem ha bandeira com as armas nacionaes. § <i>As bandeiras despregadas, fr. fig.</i> ; aberta, descobertamente, como quem lahe de praça rendida, e se lhe concede levar a bandeira tendida, ou desferida, despregada. § <i>Bandeira da janella</i> , a parte superior, que de ordinario se não abre. § Peça do candieiro voluvel, para cobrir a maior força da luz, que não dê nos olhos. § <i>Bandeira do milho</i> , he como huma espiga de trigo, que lhe sahe do mais alto do pé. § f. „a bandeira, por companhia, de algum official, que a tem. § f. „a bandeira da Cruz,, Arraes 3., 23. <i>Ao monte Olivete donde resplandece a bandeira da Cruz.</i> § „, <i>levantar bandeira no muro fig.</i> vencer, conseguir seu intento, como quem vai escalar praça murada. <i>Eufr.</i> 3. 2. <i>Salvo quando lhe levantardes a bandeira no muro.</i>

Pela ordem de apresentação e pela percentagem de acepções a ele dedicado, deduz-se que o militar seja o de maior saliência semasiológica nesse projeto lexicográfico, ou seja, os sentidos mais prototípicos atrelados ao verbo estariam inscritos nesse *frame*. É com esse *frame* que *bandeira* instancia a construção [X + nte], formando *bandeirante*, unidade simbólica em que a saliência se desloca metonimicamente mais uma vez, agora da companhia (significado já metonímico de *bandeira*) para a pessoa (o indivíduo vinculado à companhia).

A ideia de insígnia, soldados e companhia funciona semioticamente se compartilhados certos valores europeus, como organização política e bélica. A reboque vêm as ideias de nação, símbolo nacional, entre outros, que também são devedoras de valores europeus. É possível identificar nesse caso uma

transposição linguística, isto é, as acepções que circulam em Portugal são associados ao verbete, sem aparente influência do que se fez ou fazia linguisticamente no Brasil. Nesse dicionário, não consta o verbete *bandeirante*.

No *Dicionário da Língua Brasileira*, em cujo prólogo há um destaque para “a raridade do Dicionario em *nosso Idioma*” (Pinto 1832 – grifos nossos), as duas acepções de *bandeira* registradas não fogem à transposição (Quadro 2). Também ali não consta o verbete *bandeirante*. Isso sugere que a brasilidade pretendida no título e no prólogo parece não configurar o referencial conceitual do dicionário. A significação do verbete se apoia em *frames* preponderantemente lusitanos. Até aqui não se recupera na palavra outra memória se não aquela construída alhures pelo não brasileiro. O legado cultural nela semiotizado não sinaliza nenhum aspecto do que se historicizava no Brasil.

**Quadro 2. Bandeira/Bandeirante: Lexicografia luso-brasileira séc. XIX**

Elementos macroestruturais	Elementos microestruturais	
Obra	Verbetes	Acepções
<i>Dicionario da Língua Brasileira</i> (1832)	Bandeira, s. f.	Insígnia militar. Insígnia de navio.
<i>Vocabulário Brasileiro</i> (1853)	Bandeira	um indeterminado numero de homens, que providos de armas, munições, e mantimentos necessários para sua subsistência e defeza, entram nas matas virgens com o intuito de descobrir minas, reconhecer o paiz, ou castigar os selvagens, que assaltam as propriedades rurais e os viajantes, ou ainda para os civilisar.
	Bandeirante ou bandeirista	indivíduo que pertence à bandeira.

Já no *Vocabulário Brasileiro*, que, como o próprio subtítulo aponta, serve de complemento aos dicionários da língua portuguesa, destacam-se acepções que revelam aspectos de relações que contam um trecho da história do Brasil (Quadro 2). O *frame* militar bem como os valores lusitanos não são deixados de lado, embora sejam ajustados quanto à saliência. A ideia de grupo de homens armados reforça que o referencial bélico não mudou. Não há invocação da instituição militar propriamente dita, ainda que também não haja indicação de que se trata de ajuntamento armado de homens de natureza

particular, privada. Entretanto, suas atividades e a referência espacial indicam certa acomodação conceitual, especialmente no que tange à extensão.

As “matas virgens” identificam as terras brasileiras como local onde “homens providos de armas e munições” desempenham suas atividades, cujos implícitos confirmam a manutenção da perspectiva lusitana: “descobrir minas, reconhecer o paiz, ou castigar os selvagens, que assaltam as propriedades rurais e os viajantes, ou ainda para os civilizar” (Quadro 2). O conceito de “país” alinha o ato responsável do verbete a um universo cultural europeu, dado que o autóctone tinha outros modelos de sociação. Do mesmo modo, é da Europa que vem a distinção conceitual de “selvagem” e “civilizado” e, portanto, somente nesse universo simbólico faz sentido “castigar selvagens” e/ou “civilizá-los”.

Se o referencial conceitual se mantém marcadamente lusitano, as relações deflagradas aqui constroem tal registro lexicográfico, sob pena de se perder da memória da língua algo que fugia ao repertório propriamente lusitano. Mantém-se o referencial conceitual, mas altera-se ideologicamente o *frame*, com a inclusão na acepção de referentes específicos das relações entretecidas na colônia. O militar propriamente dito não constitui mais o destaque, e sim organização de homens armados e com munição; o *tópos* até então não mencionado passa a integrar explicitamente o fundo. Não se trata mais do pertencimento a uma companhia simbolizada por uma insígnia militar, e sim do papel ativo em outras terras (Quadro 3). Referenciais lusitanos transformados e não meramente transpostos fundamentam a produção do verbete e corroboram um legado cultural diferente daquele propriamente lusitano.

Saltando desses primeiros para registros mais recentes, no século XXI, evidencia-se o rumo dessa trajetória léxico-semântica cujas pistas foram deixadas nos séculos XVIII e XIX. No *Grande Dicionário Houaiss da Língua Portuguesa*, as acepções que retomam os referenciais cotejados registram: “história militar/Portugal”, datado de 1526 e 1626, e “por extensão, história/Brasil”, datado de 1698 (Quadro 3). A distinção de lugar é essencial para interpretar o que se consolida na memória da língua documentada lexicograficamente e, assim, se estabelece como legado cultural partícipe das sociações que constroem a brasilidade. A menção a Portugal resgata a gênese conceitual e opera como referência para a acepção que se constrói ‘por extensão’ e que é registrada pouco tempo depois, no mesmo século. A datação dessas acepções marca a cronologia de diferentes repertórios a partir dos quais é possível fazer sentido de tais verbetes. A memória da língua se constrói então pela distinção de *topoi* que produzem diferentes legados ativados por *frames* já marcadamente distintos. Assim, mesmo guardando um referencial conceitual lusitano, o impacto das relações socioculturais e, portanto, da

ideologia tal como compreendida por Voloshinov (2013), se não promove, certamente corrobora diferentes memórias: uma lusitana propriamente dita, ádvena, e uma brasileira, vernácula.

**Quadro 3. Bandeira/Bandeirante: Lexicografia brasileira séc. XXI**

Elementos macroestruturais	Elementos microestruturais	
Obra	Verbetes	Acepções
<i>Grande Dicionário Houaiss da Língua Portuguesa</i> (2012)	Bandeira	4 (1526) hist.mil; <i>P</i> pequeno grupo armado, de militares 4.1 (1626) hist.mil; <i>P</i> na legislação militar portuguesa consolidada por D. Sebastião (1554-1578), unidade militar sob o comando de um capitão e correspondente à companhia 5 (c1698) <i>p.ext.</i> ; hist; <i>B</i> cada uma de uma série de expedições, particulares ou oficiais, de penetração do território brasileiro na época colonial (sXVI a XVIII), que ger. partia da capitania de São Vicente (atual São Paulo SP) e tinha como objetivos fundamentais a captura de indígena e a detecção de jazidas de pedras e metais preciosos [As bandeiras foram responsáveis pelo alargamento do território brasileiro, pois ger. não respeitavam os limites impostos pelo Tratado de Tordesilhas.] 6 (c1698) hist, rel; <i>B</i> associação de escravos e ex-escravos em grupos, de acordo com seus ofícios, tendo um santo católico como padroeiro 7 rel; <i>B</i> cortejo em homenagem a santos, em cuja frente se carrega uma bandeira ou estandarte com a imagem do santo, realizado ger. em zonas rurais e em cidades pequenas ao som de instrumentos e cantos
	Bandeirante Substantivo masculino (1817) B	hist indivíduo que no Brasil colonial tomou parte em <i>bandeira</i> ('expedição'); bandeirista, bandeireiro
	Substantivo feminino	menina ou mulher que pertence à Federação de Bandeirantes do Brasil, ou que se dedica ao bandeirantismo
	adjetivo e substantivo de dois gêneros	1 p.ext. m.q. paulista ('natural ou habitante') 2 p.metf. que ou o que abre caminho; desbravador, precursor, pioneiro
	Adjetivo de dois gêneros (1871)	3 próprio de bandeirante (em todas as acp.) 4 relativo ao bandeirantismo

Atentando para a acepção 6 no *Grande Dicionário Houaiss da Língua Portuguesa* (Quadro 3), é possível notar que, além da distinção de *topoi*, houve abertura de novo *frame*, o religioso, também anotado como brasileiro. Essa abertura é promovida por outra incidência da metonímia do mesmo tipo: antes, o grupo de europeus pertencentes à companhia em Portugal; agora, o grupo de escravos associados ao mesmo ofício ou crença. O destaque aí fica por conta do ajuntamento de pessoas em torno de uma bandeira, saindo o aspecto militar bélico. Curiosamente, refere-se ao ajuntamento de escravo e ex-escravos em torno de um santo católico. Pela datação (1698), como não há menção a nenhuma etnia, supõe-se que sejam indígenas não mais escravizados, uma vez que a escravidão de africanos se manteria até final do século XIX. Independentemente disso, se é estranha a associação em torno de santos católicos, *frame* absolutamente lusitano e adventício para autóctones e africanos mas já em mescla com o *frame* religioso africano por analogia (cf. Fauconnier & Turner 1998; 2002), não é estranha a incidência do mesmo processo cognitivo (metonímia).

A despeito desse estranhamento histórico, ressalta-se a mudança semântica registrada como ‘brasileirismo’: o destaque deixa de ser *ajuntamento de homens armados sob o comando de um comandante e identificados pela insígnia/bandeira* para *ajuntamento de pessoas em torno de determinada bandeira*, agora símbolo religioso. Essa perspectiva é própria das relações estabelecidas na colônia. Pela anotação, deduz-se que, em Portugal, a unidade simbólica *bandeira* não ativa o *frame* religioso, sendo esta uma ativação local, embora ainda com base em referenciais de lá.

A acepção 7, cuja datação não é indicada (Quadro 3), é ainda mais abrangente ao retirar de destaque a condição de [ex-]escravos para a de integrantes do ajuntamento religioso em torno de uma bandeira. Se comparada com a noção de “procissão”, esta claramente lusitana, a acepção 7 de *bandeira* pode uma evidência de um tipo de saliência onomasiológica, qual seja, a *prevalência sociolinguística* (Geeraerts 2006c, p. 90), isto é, uma forma linguística que, em comparação com outras, identifica preferência em dada variedade específica da língua – no caso, português brasileiro – ou em dado contexto pragmático – no caso, *tópos* brasileiro.

O registro hodierno dos verbetes pelo instrumento produzido pela Porto Editora parece guardar apenas parte do que se refrata no Brasil (Quadro 4). No *Dicionário infopédia da Língua Portuguesa com Acordo Ortográfico*, também há a anotação “história”, mas conta-se apenas parte do que diz respeito ao *tópos* brasileiro, sem revelar a gênese que está na organização militar lusitana, nem indicar que o que ali se apresenta configurou,

ao menos em algum momento, uma “extensão de sentido”. No entanto, o registro do *frame* militar, o original, impede o apagamento da trajetória histórico-semântica da unidade simbólica.

**Quadro 4. Bandeira/Bandeirante: Lexicografia lusitana séc. XXI**

Elementos macroestruturais	Elementos microestruturais	
Obra	Verbetes	Acepções
<i>Dicionário infopédia da Língua Portuguesa com Acordo Ortográfico (2003-2017)</i>	Bandeira	HISTÓRIA expedição armada que antigamente ia explorar os sertões do Brasil
	Bandeirante adjetivo de 2 géneros, nome de 2 géneros	HISTÓRIA diz-se de ou indivíduo pertencente a uma bandeira (expedição armada) que ia explorar o sertão brasileiro

Essa distinção de memórias produzidas pelos repertórios construídos no Brasil e em Portugal impacta ideologicamente o *frame* militar acionado pelos projetos lexicográficos. Isso materializa uma divergência semântica entre o português brasileiro e o europeu e indicia parte da matriz vernácula brasileira, que guarda a perspectiva de outrem para conceituar relações próprias da(s) sociedade(s) que se estabelecia(m) na América. No caso de *bandeira*, aquilo que se registra como “brasileirismo” constitui uma conceituação lusitana para relações sociais travadas na colônia, que, por distar das experiências culturais portuguesas na Europa, reorganiza as conceptualizações a ponto de, por um lado, haver formação de nova palavra para conceituar determinado modo de participar na história colonial brasileira (ver discussão na sequência) e, por outro, diluir, em importante projeto lexicográfico português, a gênese ideológica dos vocábulos.

Diferente de *bandeira*, *bandeirante* é registrado como *brasileirismo* desde a gênese, ainda que tal anotação só apareça posteriormente na lexicografia. É *brasileirismo* não por capturar um modo propriamente brasileiro de ver as relações sociais implicadas no conceito que indicia, mas pela função comunicativa de designar acontecimentos próprios da história do Brasil. Que pistas linguístico-cognitivas são aí encontradas?

Uma possível pista é morfológica, que leva à segunda exigência reconhecida acima, a de tratar das construções gramaticais envolvidas. Nos termos da Gramática Cognitiva, o quadro 2, ao informar sobre a criação da palavra *bandeirante*, informa sobre a existência da construção [X + nte], já mencionada. O uso criativo de uma construção lhe confere o *status* de molde



(Kewitz, Almeida & Souza 2018), o que, por sua vez, depende de que a construção tenha sido depreendida de palavras anteriores (*cf.* Basilio 2010).

O mesmo Quadro 2 também apresenta o convívio entre os sufixos *-nte* e *-ista*, que, para Basilio (1995; 2008), estão a serviço da formação de agentivos no português brasileiro. Sem reproduzir aqui a alentada descrição formulada pela autora, em síntese, pode-se dizer que *-nte* forma substantivos que exprimem profissões (*despachante, gerente, servente*) e adjetivos passíveis de substantivação plena (*solvente, refrigerante, detergente*), enquanto *-ista* forma substantivos que exprimem atuação (*surfista, pianista, linguista*) ou adesão (*petista, malufista, budista*). Portanto, ser exatamente o *Vocabulário Brasileiro* (1853) o que registra *bandeirante* e *bandeirista* pode ser considerado uma evidência lexicográfica da distribuição semântica entre predicação e designação, funções semânticas que causam a distinção categorial entre adjetivo e substantivo.

O Quadro 3 acrescenta *bandeireiro* à rede de construções. Para Basilio (1995), a distinção entre atuação e adesão quanto ao sufixo *-ista* tem como uma das consequências descritivas identificar que somente ao indicar adesão *-ista* se relaciona com *-ismo* (*petista/petismo, malufista/malufismo, budista/budismo*) e que somente ao indicar atuação *-ista* se relaciona com *-eiro*, caso em que a expressão por *-ista* indica mais prestígio social (*jornalista, pianista*) e a expressão por *-eiro* indica menos prestígio social (*jornaleiro, planeiro*), mesmo quando as formações não se dão a partir da mesma base (*motociclista/motoqueiro*). A pejoração seria expressa por *-eiro* – diacronicamente, a evolução de *-arium* > *-ario* > *-airo* > *-eiro* (razão pela qual convivem formações herdadas por evolução natural (*primeiro*) e formações herdadas por evolução erudita (*primário*) – pejoração que pode ser o fator de distinção de *bandeireiro* em relação a *bandeirante* e *bandeirista* e motivação da própria designação *brasileiro* (único gentílico em *-eiro* em português).

A complexidade introduzida pelo quadro 3 corrobora a necessidade de contemplar tanto o caráter distribuído da cognição, de modo a descrever e explicar conceptualizações lectais, como pretende a Sociolinguística Cognitiva, quanto o caráter situado da cognição, de modo a identificar “(...) uma perspectiva baseada em nosso conhecimento, crença e atitudes tanto quanto em nossa posição espaço-temporal” (Croft & Cruse 2004, p. 58 – tradução nossa), como pretende a Linguística Cognitiva, entendendo-se cognição como constituída por língua e cultura, como defende a Gramática Cognitiva.

Preservadas as funções semânticas identificadas por Basilio (1995) de um ponto de vista gerativo com forte pendor ao fator semântico, abandonada a abordagem por regras (onde cabem as noções de base e derivação) e adotada a abordagem construcional segundo a Gramática Cognitiva, as funções semânticas assumem o *status* de motivação para a distribuição entre as formas, e três construções podem ser identificadas: [X + *nte*], [X + *ista*] e [X + *eiro*], das quais se toma a primeira como molde para a formação de *bandeirante*.

A instanciación desse molde por um substantivo, por sua vez, remete à formação de nomes de agente denominais (cf. Basilio 2004), como *cadeirante* e *calmante*, diferentemente da formação de nomes de agente deverbais, como os mencionados acima. Não somente existe o padrão de formação de nomes de agente denominais na língua, a exemplo de *lixeiro* e *jornaleiro*, como não há plausibilidade em aventar um verbo *cadeirar* ou *calmar*, embora seja fato que o verbo *acalmar* apresente a variante *calmar*. Se *cadeirante* é quem se move por meio de cadeira de rodas e *calmante* é a substância que produz calma (subst.. masc.), não a que calma (3ª p. sing. de *calmar*), consideram-se *bandeirante*, *calmante* e *cadeirante* como instâncias da construção [X + *nte*], especificada como [S + *nte*], distinta de [V + *nte*].

Convém observar que o *Dicionário infopédia da Língua Portuguesa com Acordo Ortográfico* da Porto Editora registra *cadeirante* também como brasileirismo e que o *Grande Dicionário Houaiss da Língua Portuguesa* descreve *cadeirante* como *cadeirar* + *nte*. A classificação como brasileirismo remete ao primeiro objetivo deste artigo, relativo às condições cognitivo-discursivas que orientam os registros lexicográficos do português brasileiro.

A abordagem construcional aqui assumida permite reconhecer que (i) –nte atua na formação de agentivos denominais (*bandeirante*, *cadeirante*), assim como –eiro (*sapateiro*, *livreiro*); (ii) –nte atua na formação de agentivos deverbais (*despachante*, *gerente*), assim como –ista (*linguista*, *projetista*); e que construções podem ser instanciadas não somente por palavras, exatamente porque o requisito é ser uma unidade simbólica, não determinada configuração estrutural. Isso sugere, novamente, uma caracterização do brasileirismo pela perspectiva estrangeira. Nesse caso, o português europeu apresenta apenas a construção [V + *nte*]. Como o português brasileiro apresenta, além da construção [V + *nte*], formações que permitem aventar a construção [S + *nte*], as formações são descritas, outra vez mais, a partir da perspectiva estrangeira. Constituem brasileirismos, mas fazendo caber na construção compartilhada com o português europeu.

## 4. Conclusão

Neste trabalho, verifica-se o que registros lexicográficos, na condição de atos responsáveis, contam do processo de historicização que se dá como legado cultural pela memória da língua portuguesa no Brasil. Para isso, foram traçados dois objetivos específicos: 1) identificar condições cognitivo-discursivas próprias do português brasileiro que impactam seu estatuto vernáculo e 2) descrever um caso de mudança semântica que ilustra novas conceptualizações no léxico interno e registros do léxico externo a serviço de um projeto lexicográfico caracterizado como ato responsável.

Quanto ao primeiro, *bandeira* e *bandeirante* foram tomados como verbetes. Desde o século XVIII, verificou-se que não há substituição de um referencial conceitual lusitano por um referencial conceitual brasileiro. Entretanto, a inserção de *frames* acompanhando novas conceptualizações associadas aos itens lexicais analisados produz uma memória distinta da língua em Portugal. Essa distinção de memórias de que se vale a política lusófona do Brasil caracteriza o fenômeno como *do* Brasil, ainda que por meio de uma cicatriz simbólica marcada pela perspectiva que parece se manter lusitana. Cicatriz simbólica porque o olhar de outrem constitui o lugar de onde se conceitua, embora, na atualidade, esse outrem não necessariamente reconheça esse olhar como seu. Trata-se de um lugar de alteridade aparentemente nunca reivindicado como de identidade. E é justamente aí que parece se desenhar cognitivo-discursivamente o vernáculo brasileiro.

Quanto ao segundo objetivo específico, *bandeira* e *bandeirantes* foram considerados como unidades simbólicas, instanciações de construções gramaticais. Para este objetivo, as pistas fornecidas pelos dicionários permitem acompanhar a mudança semântica e o convívio com outras unidades simbólicas; mudança e convívio constitutivos da formação do português brasileiro.

Por fim, teórico-metodologicamente, este trabalho contribui com uma proposta de integração de teorias que partem de bases epistemológicas compatíveis. Por um lado, princípios da Análise Dialógica do Discurso permitem trazer o ideológico, tão caro para a discussão em torno de manifestações de brasilidade, para o tratamento das relações léxico-semânticas destacadas. Por outro, fundamentos da Linguística Cognitiva, em geral, e da Gramática Cognitiva, em particular, permitem descrever a formação de palavras como um fenômeno semanticamente motivado.

## Referências

- Auroux, S. (1992). *A revolução tecnológica da gramatização*. Trad. de Eni P. Orlandi. Campinas: Editora da Unicamp.
- Bakhtin, M. M. (2003). Apontamentos 1970–1971. *Estética da criação verbal* (pp. 367–392). Trad. de Paulo Bezerra (4ª ed.) São Paulo: Martins Fontes.
- Bakhtin, M. M. (2010a). *Problemas da poética de Dostoiévski*. Trad. de Paulo Bezerra (5ª ed.) Rio de Janeiro: Forense Universitária.
- Bakhtin, M. M. (2010b). Formas de Tempo e de Cronotopo no Romance. Ensaios de poética histórica. In A. F. Bernardini et al. (Trans.), *Questões de literatura e de estética. A teoria do romance* (pp. 211–362). (6ª ed.) São Paulo: Hucitec Editora.
- Bakhtin, M. M. (2010c). *Para uma filosofia do ato responsável*. São Carlos: Pedro & João Editores.
- Barros, D. L. P. (2000). O discurso do dicionário. *Alfa*, 1(44), 75–96.
- Basilio, M. (1995). O fator semântico na flutuação substantivo/adjetivo em português. In J. Heye (Ed.), *Flores verbais*. Rio de Janeiro: Editora 34.
- Basilio, M. (2004). *Formação e classes de palavras no português do Brasil*. São Paulo: Contexto.
- Basilio, M. (2008). Substantivação plena e substantivação precária. In C. A. Gonçalves; M. L. L. de Almeida (Eds.), *Diadorim*, 4(1), 11–24. Rio de Janeiro: UFRJ
- Basilio, M. (2010). Abordagem gerativa e abordagem cognitiva na formação de palavras: considerações preliminares. *Linguística*, 6(2), 1–14.
- Bluteau, D. R. (1789). *Diccionario da Lingua Portuguesa*. Reformado e acrescentado por Antonio de Moraes Silva. Tomo I A-K. Lisboa: Officina de Simão Thaddeo Ferreira.
- Biderman, M. T. C. (2003). Dicionários do português: Da tradição à contemporaneidade. *Alfa*, 47(1), 53–69.
- Brait, B. & Magalhães, A. S. (Orgs.) (2014). *Dialogismo: Teoria e(m) prática*. São Paulo: Terracota Editora.
- Croft, W. & Cruse, A. (2004). *Cognitive linguistics*. Cambridge: Cambridge University Press.
- Dancygier, B. & Sweetser, E. (2014). *Figurative language*. Cambridge: Cambridge University Press.
- Dias, L. F. (1996). *Os sentidos do idioma nacional: As bases enunciativas do nacionalismo linguístico no Brasil*. Campinas: Editora Pontes.
- Fauconnier, G. & Turner, M. (1998) Conceptual integration networks. *Cognitive Science*, 22(2), 133–187.
- Fauconnier, G. & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. Nova York: Basic Books.
- Fillmore, C. J. (1977a). The case for case reopened. In P. Cole & J. Sadock (Ed.), *Syntax and semantics 8: Grammatical relations*. Nova York: Academic Press.

- Fillmore, C. J. (1977b). Scenes-and-frames semantics. In A. Zampolli (Ed.), *Linguistic structures processing*. Amsterdão & Nova York: North-Holland Publishing Company.
- Fillmore, C. J. (1982). Frame semantics. In The Linguistic Society of Korea (Ed.), *Linguistics in the morning calm*. Seoul: Hanshin.
- Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2), 222–254.
- Fillmore, C. J. & Atkins, B. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. In A. Lehrer & E. F. Kittay (Eds.), *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization* (pp. 75–102) Hillsdale, NJ: Lawrence Erlbaum Associates.
- Fiorin, J. L. (2009). A construção da identidade nacional brasileira. *Bakhtiniana. Revista de Estudos do Discurso*, 1(1), 115–126.
- Geeraerts, D. (2006a). The lexicographical treatment of prototypical polysemy. In *id.*, *Words and other wonders: papers on lexical and semantic topics. Cognitive Linguistics Research*, vol. 33 (pp. 327–344), Mouton de Gruyter.
- Geeraerts, D. (2006b). The definitional practice of dictionaries and the cognitive semantic conception of polysemy. In *id.*, *Words and other wonders: papers on lexical and semantic topics. Cognitive Linguistics Research*, vol. 33 (pp. 345–366), Mouton de Gruyter.
- Geeraerts, D. (2006c). Salience phenomena in the lexicon. A typology. In *id.*, *Words and other wonders: papers on lexical and semantic topics. Cognitive Linguistics Research*, vol. 33 (pp. 74–98), Mouton de Gruyter.
- Kewitz, V., Almeida, M. L. L. de & Souza, J. L. L. de. (2018). Preposições complexas: modos e moldes. In A. Tenuta & S. Coelho (Eds.), *Uma abordagem cognitiva da linguagem: perspectivas teóricas e descritivas*. Belo Horizonte: FALE/UFGM.
- Langacker, R. (1987). *Foundations of cognitive grammar*. Volume I: *Theoretical prerequisites*. Stanford: Stanford University Press.
- Langacker, R. (1991). *Foundations of cognitive grammar*. Volume II: *Descriptive application*. Stanford: Stanford University Press.
- Langacker, R. (1994). Culture, cognition, and grammar. In M. Pütz (Ed.), *Language contact and language conflict*. Amsterdam: John Benjamins.
- Langacker, R. (2008). *Cognitive grammar: a basic introduction*. Oxford: Oxford University Press.
- Langacker, R. (2009). *Investigations in cognitive grammar. Cognitive Linguistics Research*, vol. 42. Berlin: Walter de Gruyter.
- Magalhães, A. S. (2012). Políticas linguísticas e historicização do Brasil: A escrita na construção vernacular. *Gragoatá*, 17(32), 99–116.
- Magalhães, A. S. (2013). Escritas da brasilidade: Subjetivação e política lusófona na documentação vernacular. *DELTA. Documentação de Estudos em Linguística Teórica e Aplicada*, 29(1), 1–27.

- Magalhães, A. S. (2015). A palavra, os discursos e a dinâmica das memórias. *Gragoatá*, 20(32), 7–28.
- Peter, M. (2015). *Introdução à linguística africana*. São Paulo: Contexto.
- Petri, V. & Medeiros, V. (2013). Da língua partida: Nomenclatura, coleção de vocábulos e glossários brasileiros. *Letras*, 23(46), 43–66.
- Pinto, L. M. da S. (1832). Prólogo. *Diccionario da lingua brasileira*. Ouro Preto: Tipografia de Silva.
- Schneiders, C. (2017). A língua no/do Brasil: Efeitos da memória e da história. *Gragoatá*, 22(42), 329–344.
- Simmel, G. (2006). *Questões fundamentais da sociologia: Indivíduo e sociedade*. Trad. de Pedro Caldas. Rio de Janeiro: Jorge Zahar Ed.
- Silva, A. S. (2009). A sociolinguística cognitiva: Razões e escopo de uma nova área de investigação linguística. *Revista Portuguesa de Humanidades – Estudos Linguísticos*, 13(1), 191–212.
- Traugott, E. & Trousdale, G. (2013). *Contructionalization and constructional changes*. Oxford: Oxford University Press.
- Voloshinov, V. V. (2013). Que é linguagem? In *id.*, *A construção da enunciação e outros ensaios* (pp. 157–188). Trad. de João Wanderley Geraldi. São Carlos: Pedro e João Editores.

[recebido em 1 de maio de 2018 e aceite para publicação em 15 de fevereiro de 2019]



# CARACTERÍSTICAS IDENTIFICADORAS E DIFICULDADES NA APLICAÇÃO DE LISTAS PARA A ANOTAÇÃO DE ENTIDADES GEOGRÁFICAS MENCIONADAS

IDENTIFYING CHARACTERISTICS AND DIFFICULTIES WHEN USING GAZETTEERS TO ANNOTATE GEOGRAPHICAL NAMED ENTITIES

Afonso Xavier Canosa\*  
canosarodriguez@gmail.com

Na anotação automática de entidades geográficas mencionadas, as listas especializadas de topónimos têm que enfrentar ambiguidades e contextos em que o valor geográfico de uma expressão não é evidente. Neste artigo, estuda-se o caso prático de um índice de topónimos utilizado para criar um corpus anotado da *Peregrinação* de Mendes Pinto. As dificuldades achadas servem para classificar os tipos de erros que se produzem quando o topónimo é resolvido pela simples coincidência de expressões e introduzem critérios para a identificação das entidades geográficas, uma tarefa que deve preceder e tem um impacto direto nos resultados obtidos no processo de anotação automática.

**Palavras-chave:** Entidades Geográficas Mencionadas. REM. Topónimos. Anotação de corpus. Corpus histórico.

In order to annotate geographical named entities, gazetteers have to face ambiguities and contexts where the geographical value of a given expression is not clear. In this paper, an index of place names is used to examine the main problems encountered in the production of an annotated corpus of Mendes Pinto's *Pilgrimage*. The difficulties found serve to classify the types of errors that occur when the place name is solved by simple string match and introduce criteria for the identification of geographical entities, a task that should precede and has a direct impact on the results obtained in an automatic annotation approach.

**Keywords:** Geographical Named Entities. NERC. Toponyms. Corpus annotation. Historical corpus.

---

\* Universidade de Santiago de Compostela, Espanha.





## 1. Introdução

Entidade Mencionada (EM) é o termo utilizado em Processamento da Linguagem Natural (PLN) para se referir mais comumente a nomes de pessoas, organizações e lugares (Amaral *et al.* 2014; Nadeau & Sekine 2007; Santos & Cardoso 2007). O termo propriamente aponta a um referente, objeto único no mundo real, porém, é frequentemente utilizado para se referir à expressão, isto é, a cadeia de caracteres que aparece num texto para designar uma entidade.

Neste artigo, *Entidade Geográfica Mencionada* (EGM) será aquele nome que refere um objeto geográfico único (pode haver vários lugares com um mesmo nome, mas quando o usamos estamos a nos referir a um lugar em concreto e só um) instância de uma classe (refere o indivíduo e não a classe, ex. *Lisboa* é uma cidade). O exemplo mais comum são os topónimos, ainda que também pode abranger entidades menores, tais como nomes de edifícios ou ruas e, menos frequentemente, os gentílicos, enquanto se considera o seu radical como expressão portadora das características semânticas do nome próprio. Quando quiser referir o objeto geográfico, espaço físico que ocupa uma posição determinada no planeta Terra, utilizarei mais frequentemente o termo *referente*.

Um recurso comum no processo de anotação de entidades geográficas mencionadas é o uso de listas de entidades geográficas (*gazetteers*) que provêm, para além do topónimo, informação complementar de utilidade para a desambiguação e georreferenciação (Leidner 2007, p. 51; Southall, Mostern & Berman 2011), particularmente as coordenadas geográficas em termos de latitude e longitude. Na aplicação de *gazetteers* para a anotação automática, quando um termo no texto coincide com um topónimo da lista, outorgamos o atributo de entidade geográfica e recuperamos a informação relevante disponível segundo os objetivos e o problema a resolver: quer simples reconhecimento e classificação das entidades mencionadas, quer labores mais específicos de resolução e análise geográfica (Gregory *et al.* 2013; 2015). Porém, mesmo quando se tiver uma lista específica, a simples aplicação dos topónimos produz ambiguidades (ex. *Carvalho* pode ser um topónimo, antropónimo ou nome comum).

Para superar estas dificuldades, os sistemas de anotação automática introduzem listas auxiliares (de distintas categorias de EM e ativadores

da classe) e regras que permitam desfazer as ambiguidades a partir de padrões e regularidades linguísticas, do tipo: *se um nome próprio que coincide com um topónimo na lista vai precedido de antropónimo, será um apelido e não uma EGM* (ex. Paulo Carvalho) ou *se um nome próprio vai precedido de um nome comum da lista de ativadores de classe geográfica seguido pela preposição de, é uma EGM* (ex. rei de Portugal). Estas regras podem ser expressadas de modo explícito no código da ferramenta (sistema de regras) ou aprendidas a partir do treino em corpora (sistema de aprendizado de máquina). A vantagem do primeiro tipo é que, ao permitir otimizar os resultados mediante a depuração das regras, resulta facilmente adaptável para o trabalho com corpora históricos em que a modalidade de língua apresenta grandes diferenças com o padrão contemporâneo, caso do galego-português medieval (Canosa *et al.* 2018). Para serem convenientemente treinados, os sistemas de aprendizado requerem texto previamente anotado e estatisticamente relevante, num volume nem sempre disponível no caso de textos históricos. Porém, nas avaliações realizadas sobre textos em inglês dos séculos XVII e XVIII (modalidade de língua mais próxima ao padrão contemporâneo com que foram treinados) ofereceram os melhores resultados de desempenho (Won, Murrieta-Flores & Martins 2018).

## 2. Motivação e objetivos

Independentemente do sistema de anotação utilizado, o resultado final vem condicionado pelo que se entende como EGM. As regras, quer aprendidas automaticamente a partir de corpora já anotados, quer explícitas no código, necessitam de uma definição prévia do que se deve anotar. Para uma filóloga estudiosa da evolução de um topónimo, qualquer concordância serve para observar usos gráficos com que analisar possíveis alterações fonéticas, independentemente de se a expressão aparecer como apelido ou nome de lugar. Porém, para um historiador interessado em reconstruir redes de relações num momento dado, a origem toponímica de um apelido pode ser totalmente irrelevante. Há, portanto, um componente subjetivo importante à hora de definir o que deve ser uma EGM.

Doutra parte, dado que as ferramentas de que dispomos na atualidade são limitadas, faz-se necessário o trabalho em equipa para desenvolver ou melhorar as utilidades que facilitem o processo de anotação. O presente artigo pretende ilustrar as dificuldades que aparecem ao tratar de conciliar

os requerimentos próprios de quem elabora um corpus para a investigação humanística posterior com as limitações e possibilidades de automatização resultantes de aplicar uma lista que identifica as expressões coincidentes no texto (exemplo que serve de *baseline* para o desempenho de uma ferramenta de anotação mais específica).

Em textos históricos e obras comentadas com aparato crítico, é relativamente comum dispor de índices temáticos e glossários de entidades geográficas específicas ou muito próximas aos objetivos do corpus. A questão que se pretende responder aqui é, então, quais são as dificuldades que surgiram ao aplicar um destes índices para anotar automaticamente um texto numa modalidade linguística para a qual não há ferramentas específicas desenvolvidas. Que regras ou exceções necessitamos aplicar para anotar todas as entidades da lista? Quais foram os critérios utilizados para identificar uma expressão como EGM?

### 3. Materiais e métodos

Analiso a seguir as situações mais problemáticas achadas no caso prático de anotação do corpus da *Peregrinação* de Fernão Mendes Pinto (1614) a partir de uma lista específica da obra. Para este corpus em particular, temos índices e dicionários (Albuquerque 1994; Alves 2010; Flores, Gomes & Sousa 1983; Lagoa 1950–1953). O mais exaustivo e específico (Alves 2010) recolhe praticamente todas as formas toponímicas, ainda que não todas as variantes, motivo pelo qual elaborei uma lista própria, extraída manualmente em sucessivas leituras que acompanharam o estudo crítico do texto. Os glossários prévios foram de todos modos muito úteis para comprovar a qualidade da lista própria e um instrumento de trabalho imprescindível para a abordagem de problemas mais avançados de resolução geográfica.

Mediante uma série de scripts que processam o corpus, utilizei a lista própria para anotar de modo automático os topónimos no texto completo da *Peregrinação* (PR) segundo a primeira edição. O procedimento e objetivos mais específicos de georreferenciação foram publicados com anterioridade (Canosa 2017). Os resultados concretos de corpus anotado e georreferenciado estão disponíveis para consulta pública *on-line*.<sup>1</sup>

---

1 <https://www.pucau.org>

## 4. Análise e discussão das principais dificuldades detetadas na anotação do *corpus* do caso prático

Os exemplos a continuação ilustram as principais dificuldades achadas e os critérios aplicados para anotar um *token* face a outros que, ainda coincidindo na mesma expressão, não são entidade geográfica mencionada. Cada dificuldade estudada é fechada com a característica mais relevante para determinar que a expressão em questão seja ou não uma EGM. As características identificadoras servem de regras que, em conjunto, provêem a definição de entidade geográfica aplicada para o *corpus*. A referência às concordâncias que ilustram os casos são feitas para o capítulo de modo que sejam recuperáveis em qualquer edição.

### 4.1. A entidade geográfica mencionada coincide com uma forma do vocabulário

Sejam as concordâncias:

- a) “Como Antonio de Faria chegou ao rio de Tinacoreu, a que os nossos chamão **Varella**, & da informação que daquelle reyno lhe derão hūs mercadores.” (PR, 41)
- b) “& no mesmo dia se coroou por Rey de Péguu **na varella grande**” (PR, 193)
- c) “Passados os dez dias deste encerramento, **as varellas** & pagodes, & brallas, que são os seus templos, amanheceraõ todos ornados de insignias de alegria” (PR, 184)

A forma normalizada “varela” tem vários significados. No contexto da *Peregrinação* achamos o recolhido por Pereira (1647) e Bluteau (1712–28):

“Varela. Templum” *Thesouro*, (Pereira 1647, fól. 94 v.)

“Varella, ou Varela. (Termo da India.) Templo de Idolos, ou mosteyro de Gentios.”  
*Vocabulario*, (Bluteau 1712–28)

Estamos, portanto, perante um nome comum que se regista como tal num dicionário. No entanto, na primeira cita (a) achamos a forma precedida

por um topónimo, *Tinacoreu*, explicitando ademais o seu uso como nome de lugar, a forma portuguesa equivalente a uma outra asiática. Isto é, temos um topónimo transparente, um nome de lugar com um significado que se corresponde, para além da denotação de um espaço geográfico particular, com o de um nome comum, mas nome de lugar nesta concordância e, assim sendo, EGM.

Característica identificadora: a entidade geográfica mencionada começa por maiúscula. Assim (a) é EGM, (b) e (c), sendo a mesma expressão, não são entidades geográficas mencionadas.

#### 4.2. A entidade geográfica mencionada coincide plenamente com uma forma do vocabulário

Uma dificuldade maior aparece na seguinte concordância:

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

Citamos a forma transparente *a varella* como equivalente a um topónimo opaco, *Tinacoreu*, mas agora com uma particularidade, não se faz uso de maiúscula. No entanto o seu valor toponímico é tão claro como o citado anteriormente em (a). Temos um caso de variação ou dúvida na normalização, o editor mostra incoerência ou houve gralha na publicação, característica dos textos históricos, face ao processamento de textos contemporâneos em que aguardamos aderência a uma norma que nos permita sistematizar todos os casos e as suas exceções. Em (d) temos uma EGM, mas a sua codificação supõe uma exceção à primeira regra de identificação: não começa por maiúscula. No nosso caso anotamo-la igualmente como EGM e a lista regista-a como mais uma variante com a particularidade de aparecer em minúscula.

Seguindo o mesmo critério anotamos como EGM:

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

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

- f) “sem fazeres nenhũa detença te venhas logo com essas naos por junto do **baluarte do caez da varella**, onde me acharàs em pé esperâdo por ty” (PR, 148)
- g) “E assi no tempo que o Rey Bramaa foy sobre o reyno de Sião, & após cerco à cidade de Odiaa, como atras fica dito, pregando o Xemindoo então **na varella do Comquiay de Pegù**, que he como See de todas as outras” (PR, 190)
- h) “E com isto se partio logo para a cidade de Pegù, onde dos moradores della foy recebido com triumpho de Rey, & coroado por esse **na varella do Comquiay**, que he como See de todas as outras.” (PR, 190)

Mesmo se todas as frases destacadas em (d), (e), (f), (g) e (h) referem um lugar concreto, que pode ser referenciado (e nesse sentido todas cinco são georreferências susceptíveis de lhe serem atribuídas umas coordenadas geográficas específicas), em (d) e (e) *a varella* é instância da classe *rio*, e não da classe *varela* (templo). Tem ademais o valor de unicidade, apenas há um *rio da Varela*. Propriedades, estas, de objeto único e indivíduo (e não classe) apontadas na introdução (§1) como características mais definitórias da EM.

No entanto em (f), (g), (h), a mesma expressão refere tanto o indivíduo quanto a classe (*varela*). Precisa, ademais, de modificadores para denotar uma individualidade (“*do Comquiay*” (g), (h)), ou aparece simplesmente como modificador duma individualidade (“*do baluarte do caez da varella*” (f)).

*Característica identificadora: a entidade geográfica mencionada aparece em contexto de nome próprio.* Assim em (d) e (e) temos nomes próprios que grafariamos com maiúscula segundo as convenções atuais, no entanto em (f), (g) e (h) estamos perante nomes comuns. Para a identificação de (d) e (e) como EGM necessitamos, portanto, criar uma exceção à regra das maiúsculas (§4.2). As marcas identificadoras são agora os elementos lexicais que precedem a EGM (verbo *chamar* e termo geográfico *rio*).

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

Seja a concordância:

- i) “Como nos partimos desta **ilha dos ladroões** para o porto de Liampoo, & do que passamos atê chegarmos a hum rio que se dizia Xingrau” (PR, 55)

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

- j1) “& com este concerto jurado & assinado por todos, se vieraõ surgir a hũa **ilha que se dizia dos ladroës**” (PR, 53)

o mesmo topónimo aparece com uma cláusula intercalada. Pelo critério usado em §4.2 para (e), (i) é também EGM, mas em (j) *ilha* comporta-se como um nome comum (de facto vem precedido do artigo indefinido para a singularizar, enfatizando o significado de não unicidade do termo). Consideramos três opções:

- 1) Simplificamos o topónimo e anotamos *dos ladroës* como variante.

De um ponto de vista estritamente linguístico, esta é a solução mais coerente com o estatuto gramatical das EGMs se quisermos manter uma cadeia única e contínua (não interrompida). Porém, da parte do processamento automático, surge mais uma dificuldade: ao aplicarmos a lista sobre o corpus obtemos também a concordância:

- k) “mãdou tambem hum Naique com vinte Abexins que nos veyo guardando **dos ladroës**, & prouendonos de mâtimêto & caualgaduras ate o porto de Arquico” (PR, 4)

Uma possível solução à ambiguidade criada por concordâncias como (k) requer processar não só as entidades geográficas mencionadas, mas o contexto em que aparecem. A anotação morfossintática mediante técnicas de PLN identifica um verbo (*guardando*) antes da frase preposicional (*dos ladroës*), uma regra simples indica que neste caso não se trata de uma EGM.

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

- j2) “& com este concerto jurado & assinado por todos, se vieraõ surgir a hũa <LUGAR>**ilha que se dizia dos ladroës**</LUGAR>” (PR, 53)

3) Uma solução intermédia adiciona um módulo no sistema de aplicação da lista ao corpus que identifica o início do topónimo (*ilha*) e resto dos componentes (*dos ladroës*), obviando os elementos alheios (*que se dizia*). A dificuldade desta proposta é termos de processar a lista de variantes de entidades geográficas identificando os seus componentes. Neste exemplo, a marca <B> assinala o começo da entidade mencionada, <O> o segmento a omitir e <I> a parte final.

- j3) “& com este concerto jurado & assinado por todos, se vieraõ surgir a hũa  
 <LUGAR><B>ilha<B> <O>que se dezia</O> </I>dos ladroẽs</I></  
 LUGAR>” (PR, 53)

Optamos pela solução 2, operativamente mais eficaz no processamento do corpus, já que requer unicamente adicionar mais uma variante na lista. As soluções alternativas guardam uma maior homogeneidade nas variantes do lexema mas dificultam o processamento, ao necessitarem de uma regra para solucionar um único caso no corpus.

*Característica identificadora: a entidade geográfica mencionada incorpora uma forma genérica de identificativo geográfico (ex. rio de, cidade de, ilha de) e mesmo sintagmas verbais adicionais quando a forma mais característica do nome próprio fica, de outro modo, incompleta.*

#### 4.4. A entidade mencionada é ambígua na expressão de referente geográfico

No exemplo:

- 11) “E sendo sua alteza certificado da sua morte, proueo segunda vez na mesma capitania a hum Diogo Cabral da ilha da **Madeyra**, a quem Martim Afonso de **Sousa** a tirou por justiça, por se dizer que praguejara delle sendo Governador, & a deu a hum Ieronymo de **Figueiredo** fidalgo do Duque de **Bargança**” (PR, 20)

temos quatro entidades geográficas mencionadas com um elemento comum: servirem de complemento a um nome próprio antropónimo, com uma função similar à dos apelidos, mas precedidas de uma preposição que permite a interpretação da frase como lugar de procedência. Cada caso apresenta alguma particularidade que não têm os outros. Aceitando a definição de entidade mencionada como portadora do princípio de unicidade (§1) estamos perante um referente de pessoa, isto é, uma solução por exemplo do tipo:

- 12) “E sendo sua alteza certificado da sua morte, proueo segunda vez na mesma capitania a hum <PESSOA> **Diogo Cabral** </PESSOA> da ilha da <LUGAR> **Madeyra** </LUGAR>, a quem <PESSOA> **Martim Afonso de Sousa** </PESSOA> a tirou por justiça, por se dizer que praguejara delle



sendo Gouvernador, & a deu a hum <PESSOA>**Ieronymo de Figueiredo**</PESSOA> fidalgo do <PESSOA>**Duque de Barchina**</PESSOA>” (PR, 20)

Importa agora notar como quatro entidades do tipo PESSOA aparecem, dentro duma estrutura sintática similar (Frase Nominal + Frase Preposicional), ligadas a entidades geográficas de modo distinto. No primeiro caso, *Diogo Cabral da ilha da Madeyra*, (óbvio agora o facto de o segundo nome próprio, *Cabral*, ser também expressão de um topónimo) a presença de um termo do domínio geográfico (ilha) evidencia que estamos perante uma entidade geográfica na frase preposicional: é indício claro de referente geográfico, a pessoa está a ser referida como procedente de um lugar concreto. No nosso corpus é anotada como EGM com a marca <LUGAR> em (I<sub>2</sub>).

No segundo caso, *Martim Afonso de Sousa*, temos uma forma (*Sousa*) apelido comum, mas também topónimo, nome de rio em N 41° 5' 48", W 8° 30' 6", onde também dá nome a freguesia, para além de ser topónimo noutras localidades de Portugal. A expressão é ambígua e temos de optar por uma interpretação. A questão a responder é, optamos por marcar Sousa como um nome de lugar ou deixamo-lo como apelido (tal e como se faz com *Cabral* em *Diogo Cabral*, independentemente de que o apelido tenha tido originalmente uma origem toponímica)? Uma solução é usarmos critérios que determinem a escolha. Considerando que a forma candidata a entidade geográfica aparece em todas as ocorrências como denotadora de pessoa, sem acharmos uso nenhum como entidade geográfica independente, optamos por deixar sem anotar como EGMs estes casos, isto é, o critério de avaliação de candidatos ao usarmos uma lista de termos geográficos penaliza a ambiguidade e favorece aquelas formas que aparecem em contextos em que o referente é mais inequivocamente geográfico.

O terceiro caso, *Ieronymo de Figueiredo*, também contém um nome de várias freguesias em Portugal. Tem a particularidade de ser transparente, isto é, leva um significado explícito associado com um morfema derivativo associado à marca de lugar (-*edo*, expressão de fitotopónimo com o significado de lugar em que abunda uma espécie). Mas sintaticamente aparece numa estrutura típica de antropónimo. Aceitando a situação de ambiguidade (um estudo filológico mais detido poderia desfazê-la com relativa facilidade) usamos os critérios de simplificação e frequência no corpus (apenas ocorre em nome de pessoa), e deixamos *Figueiredo* como mais uma expressão sem função de referente geográfico.

Finalmente temos *Duque de Barchina*. O conjunto é uma entidade mencionada de pessoa, mas agora aparece uma ligação a um referente

geográfico de forma não ambígua. Nos casos precedentes, um ou vários nomes próprios de pessoa vão seguidos de mais um nome próprio como resultado de uma codificação histórica ou cultural (por exemplo, sistemas romano, português ou inglês para o nome completo de uma pessoa) que pode, como mais uma possibilidade, ser ocupado por um nome de lugar (em função dos usos administrativos numa jurisdição, período, circunstâncias individuais mesmo). Nomes associados a estruturas governativas e territoriais requerem semanticamente um âmbito geográfico e, portanto, desaparece a ambiguidade que achamos no apelido (*de Sousa* e *de Figueiredo*). Mais ainda, em *Duque de Bargaça*, a frase nominal núcleo não contém um nome próprio, mas um comum com um significado genérico. Nestes casos sim optamos por atribuir a marca de EGM, pois o primeiro termo da entidade pessoa aponta para uma entidade geográfica de forma não ambígua. Aliás, usamos o critério de frequência, porquanto a mesma expressão tem uma ocorrência (m) em que aparece como EGM independente:

m) “hum Portuguez que andaua com elles, por nome Christouão Sarmento natural de **Bargaça**” (PR, 195)

Característica identificadora: a entidade geográfica mencionada tem como referente primeiro uma entidade geográfica.

Isto é, *Cabral* em *Diogo Cabral*, *Sousa* em *Martim Afonso de Sousa* e *Figueiredo* em *Ieronymo de Figueiredo* são, primeiramente, apelido, e aparecem no corpus unicamente como apelido: o seu valor referencial é o de um antropónimo e não o de um topónimo. Mesmo querendo atribuir-lhe um valor de nome de lugar, são expressões de grande ambiguidade geo / geo (topónimos muito comuns). Quando se quiser explicitar a procedência geográfica como modificador do antropónimo, necessitaremos mais algum elemento explicativo (*da ilha da Madeyra* em *Diogo Cabral da ilha da Madeyra*). No entanto, em *Duque de Bargaça* temos uma expressão que de modo inequívoco está a especificar um espaço geográfico, o próprio núcleo nominal requer que o modificador seja uma entidade geográfica. Enquanto o objetivo principal de anotação do corpus é unicamente o estudo da geografia (as entidades de pessoa não são anotadas), e uma boa parte dos topónimos do corpus, particularmente na Ásia, vêm precedidos por um nome comum que ativa um topónimo (termos com significado de autoridade sobre um território facilmente sistematizáveis numa lista) considera-se o referente de *Bargaça* como uma entidade geográfica e a expressão é, portanto, anotada como EGM.

#### 4.5. A expressão da entidade geográfica mencionada é uma entidade não geográfica

No exemplo:

- n) “& hum destes Portugueses era hum Christouão Doria, que nesta terra foy depois mandado por capitão a **São Tomè**, & os outros dous erão Luys Taborda, & Simão de Brito, todos homens honrados & mercadores ricos” (PR, 147)

temos uma situação inversa a §4.4, estamos perante uma EGM cuja expressão, *São Tomè*, tem também valor de hagiónimo. Neste caso o referente é claramente a entidade geográfica. Seguindo este mesmo critério, de anotarmos a expressão segundo o referente primeiro, no exemplo:

- o) “caminhamos ao longo de hum rio mais cinco legoas, até hum lugar que se chamaua Bitonto, no qual nos agasalhamos aquella noite em hum bom Mosteyro de Religiosos que se chamaua **Sao Miguel**, com muyta festa & gasalhado do Prior & Sacerdotes que nelle estauão, onde nos veyo ver hum filho do Barnagais Gouvernador deste imperio de Etyhopia.” (PR, 4)

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

Do mesmo modo o teónimo *Tinagoogoo* na concordância:

- p) “E porque o embaixador adoeceo aquy de hũ inchaço nos peitos, foy acõ-selhado que não passasse adiãte até não ser saõ delle, pelo que assentou cõ algũs dos seus de se yr curar a hũa grande enfermaria que estaua daly doze legoas adiante em hũ pagode por nome **Tinagoogoo**, que quer dizer, deos de mil deoses, para onde partio logo, & chegou là hum sabbado ja quasi noite” (PR, 158)

aparece explicitamente como expressão de um edifício, um pagode. Portanto, o seu referente é uma entidade geográfica e não diretamente o deus *Tinagoogo*, como no exemplo:

- q) “na qual noite se gastou infinito numero de cera nas luminarias que se fizeram, as quais tomauão tanto espaço de terra quanto a vista podia alcançar, o que tudo parecia então que ardia em fogo, & a razão disto era, porque

dezião que o **Tinagoogoo** deos de mil deoses era ido em busca da serpe tragadora para a matar com hũa espada que lhe viera do Ceo.” (PR, 161)

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

r) “Do caminho que fizemos até chegarmos ao pagode de **Tinagoogoo**.” (PR, 158)

A expressão *Tinagoogoo* aparece agora como modificador e não núcleo do sintagma EGM. No entanto, o contexto refere o mesmo pagode que em (p). A solução que adotamos neste caso passa por um critério alheio a regras linguísticas: com o fim de obtermos mais ocorrências para o tratamento do corpus e resolução de georreferências, anotamos como EGM aqueles casos em que, tendo sido declarada a expressão de modo explícito como entidade geográfica, volte aparecer numa outra concordância acompanhada de um atributo que permite atribuir um objeto geográfico como referente.

*Característica identificadora: uma expressão declarada de modo explícito no texto como entidade geográfica será anotada como tal sempre que mantiver o valor de referente de objeto geográfico, ainda quando tiver também ocorrências com o valor de entidade de uma classe não geográfica.*

Assim em (o) um hagiónimo é explicitamente declarado expressão geográfica, como também o teónimo em (p) nomeia de modo inequívoco um edifício, em ambos os casos anotamos a expressão como EGM. Do mesmo modo anotamos (r), porquanto ainda estando diante de um teónimo, a expressão foi declarada em mais alguma ocorrência no corpus (p) como entidade geográfica e volta aparecer como modificador do tipo geográfico que a subclassifica (pagode). Porém, em (q), a mesma expressão tem unicamente o valor de teónimo e, em consequência, não é considerada EGM.

#### 4.6. A entidade geográfica mencionada contém outra entidade geográfica mencionada

No exemplo:

s1) “& leuamos tambem hum Bispo Abexim, que vinha para vir a este reyno, & daquy yr a **Santiago de Galiza**, & a Roma, & dahy a Veneza, para dahy se passar a Ierusalem.” (PR, 5)

temos *Santiago de Galiza*, entidade mencionada interpretada como a cidade com coordenadas (N 42° 52' 49", W 8° 32' 44"), mas também *Galiza*, entidade geográfica de âmbito maior, presente de feito no texto em gentílicos:

- t) “duas fustas em que hão sessenta Portugueses, de hũa das quais era capitão Diogo Soarez o **Galego**” (PR, 204)

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

- s2) “& leuamos tambem hum Bispo Abexim, que vinha para vir a este reyno, & daquy yr a <LUGAR>**Santiago de Galiza**</LUGAR> </LUGAR>, & a <LUGAR> Roma </LUGAR>, & dahy a <LUGAR>Veneza</LUGAR>, para dahy se passar a <LUGAR>Ierusalem</LUGAR>.” (PR, 5)

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

- s3) “& leuamos tambem hum Bispo Abexim, que vinha para vir a este reyno, & daquy yr a <LUGAR>**Santiago de Galiza**</LUGAR>, & a <LUGAR> Roma </LUGAR>, & dahy a <LUGAR>Veneza</LUGAR>, para dahy se passar a <LUGAR>Ierusalem</LUGAR>.” (PR, 5)

*Característica identificadora: a entidade geográfica mencionada tem como referente aquele que cobre o conjunto da forma complexa independentemente de um dos seus componentes ser entidade geográfica mencionada independente.*

## 5. Considerações finais

Nos exemplos analisados considerei as limitações surgidas ao aplicar uma lista que contém todas as entidades geográficas mencionadas no corpus num processo de anotação por coincidência de expressões. Aachamos problemas que resultam da dificuldade para determinar se os *tokens* recuperados são EGMs ou não. A listagem não pretende ser exaustiva, mas simplesmente servir de mostra das ambiguidades mais comuns que aparecem ao tentar automatizar o processo de anotação. Embora o caso prático aqui estudado atende as EGM, é de esperar que problemas similares aconteçam com outro tipo de entidades, como ficou parcialmente ilustrado nos exemplos de ambiguidades relativas a nomes de pessoas.

Para cada dificuldade aponte como conclusão uma característica identificadora, sem que isto implique que seja necessariamente a única solução possível. Os critérios aplicados evidenciam a subjetividade inicial da anotação que, necessariamente, condicionará o desempenho de uma ferramenta mais avançada, particularmente quando esta não for concebida nem adaptada para os objetivos específicos do corpus.

As expectativas criadas na automatização da anotação de entidades mencionadas devem, portanto, considerar que uma parte importante do problema de automatização reside na definição prévia do que se deve ou não anotar. Uma mesma ferramenta, mesmo em supostos muito favoráveis, em que trabalha com uma lista com abrangência total das entidades contidas no texto, como é o caso aqui analisado, terá de resolver ambiguidades cuja solução será satisfatória em função dos critérios previamente definidos por quem vai operar com o produto final, o corpus anotado.

Os exemplos mostram também como as regras necessárias para a identificação de uma entidade como EGM chegam a ser de uma especificidade tal que resulta de difícil formulação em termos morfosintáticos: a solução passa mais facilmente por um critério semântico e os ativadores lexicais, de os querer utilizar, obrigam à consideração de regras muito específicas. Se se operar por treino, os casos são tão particulares que dificilmente têm relevância estatística (não há concordâncias suficientes). O trabalho da especialista é que determina o necessário equilíbrio entre a validação experta e a adequação e melhora da ferramenta de automatização, imperfeita, porém, suficientemente eficaz para fazer desnecessária uma boa parte do tedioso trabalho de anotação manual.

## Agradecimentos

Os parágrafos da secção 4 deste artigo foram inicialmente redigidos como parte da tese de doutoramento defendida na Universidade de Santiago de Compostela intitulada *Entidades geográficas mencionadas. O caso da Peregrinação de Fernão Mendes Pinto*. O autor agradece a orientação e comentários dos professores Paulo Gamallo, Rubén Lois González e José António Souto durante o período de redação da tese. A ideia de dar forma de artigo surgiu pelo convite realizado pelo professor Xavier Varela, também da USC, para participar no congresso *Lingüística Histórica e Toponímia Galego-Portuguesa* celebrado em Santiago de Compostela, do 25 ao 26 de janeiro de 2018. A preparação do texto final em forma de artigo foi realizada

durante o período de trabalho no projeto de desenvolvimento de uma ferramenta para a anotação de topónimos em textos medievais no CITIUS da USC (Outubro 2017 – Fevereiro 2018) no marco da rede galega de investigação TECANDALI, ED341D R2016/011. Finalmente, o autor agradece os comentários e sugestões do professor Alberto Simões da Universidade do Minho e os pareceres recebidos no processo de avaliação para este volume da *Diacrítica* que contribuíram para uma nova versão, notavelmente acrescentada, do artigo.

## Referências

- Amaral, D. O., Fonseca, E. B., Lopes, L. & Vieira, R. (2014). Comparative Analysis of Portuguese Named Entities Recognition Tools. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik (pp. 2554–2558). European Language Resources Association (ELRA). Disponível em: <[http://www.lrecconf.org/proceedings/lrec2014/pdf/513\\_Paper.pdf](http://www.lrecconf.org/proceedings/lrec2014/pdf/513_Paper.pdf)>.
- Canosa, A. X. (2017). Algumas interseções disciplinares na recuperação da geografia da *Peregrinação* de Fernão Mendes Pinto. *Fluxos e Riscos*, 2(1).
- Canosa, A. X., Varela, X., Lema, P., Gamallo, P., Taboada, J. A. & Garcia, M. (2018). Uma utilidade para o reconhecimento de topónimos em documentos medievais. *Linguamática*, 11(1).
- Gregory, I. N., Baron, A., Murrieta-Flores, P., Hardie, A. & Rayson, P. (2013). Geographical Text Analysis Mapping and spatially analysing corpora. In A. Hardie, & R. Love (Eds.), *Corpus Linguistics 2013 Abstracts* (pp. 105–108). UCREL. Disponível em: <<http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf>>.
- Gregory, I. N., Cooper, D. C., Hardie, A. & Rayson, P. (2015). Spatializing and Analyzing Digital Texts: Corpora, GIS, and Places. In D. J. Bodenhamer, J. Corrigan, and T. M. Harris (Eds.), *Deep Maps and Spatial Narratives*. Bloomington: Indiana University Press. Disponível em: <<http://e-space.mmu.ac.uk/579357/2/Spatializing%20and%20Analyzing%20Digital%20Texts.pdf>>.
- Leidner, J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names* (Tese de PhD, University of Edinburgh,). Disponível em: <<https://www.era.lib.ed.ac.uk/handle/1842/1849>>.
- Nadeau, D. & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), 3–26. Disponível em: <<http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>>.

- Santos, D. & Cardoso, N. (Eds.). (2007). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área*. Linguatca 2007. Disponível em: <<http://comum.rcaap.pt/bitstream/10400.26/380/1/LivroSantosCardoso2007.pdf>>.
- Southall, H., Mostern, R. & Berman, M. L. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5(2), 127–145.
- Won, M., Murrieta-Flores, P. & Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5(2). doi: <https://doi.org/10.3389/fdigh.2018.00002>

## Fontes e estudos para a lista de topónimos e corpus

- Albuquerque, L. (Dir.). (1994). *Dicionário de História dos Descobrimentos Portugueses*. 2 vols. Lisboa: Caminho.
- Alves, J. S. (Dir.). (2010). *Fernão Mendes Pinto and the Peregrinação*. 4 vols. Lisboa: Fundação Oriente.
- Bluteau, R. C. R. (1712–28). *Vocabulario portuguez e latino, aulico, anatomico, architectonico, bellico, botanico, brasílico, comico, critico, chimico, dogmatico, dialectico, dendrologico, ecclesiastico, etymologico, economico, florifero, forense, fructifero...* Coimbra, Portugal: Collegio das Artes da Companhia de Jesus. Edição digital fac-similar: Biblioteca Nacional de Portugal. Disponível em: <<http://purl.pt/13969>>.
- Flores, A. M., Gomes, R. V. & R. H. Pereira de Sousa. (1983). *Fernão Mendes Pinto. Subsídios para a sua Bio-Bibliografia*. [Almada]: Câmara Municipal da Almada.
- Lagoa, V. (1950–53). *Glossário Toponímico da Antiga Historiografia Portuguesa Ultramarina*. 4 vols. Lisboa: Junta de Investigações Coloniais.
- Pereira, B. (1647). *Thesouro da Lingoa Portuguesa*. Lisboa: Paulo Craesbecck. Edição digital facsimilar: Biblioteca Nacional de Portugal. Disponível em: <<http://purl.pt/29129>>.
- Pinto, F. M. (1614). *Peregrinaçam*. Lisboa: Pedro Crasbeek. Edição digital fac-similar: Biblioteca Nacional de Portugal. Disponível em: <<http://purl.pt/82>>.

[recebido em 7 de março de 2018 e aceite para publicação em 31 de outubro de 2018]





# UMA VERSÃO EM PORTUGUÊS EUROPEU DO C-TEST

## EUROPEAN-PORTUGUESE VERSION OF THE C-TEST

Masayuki Yamada\*

masayu36@miador.nagoya

A avaliação da proficiência é uma questão importante na interpretação dos resultados das investigações sobre a aprendizagem de língua segunda. Dados contraditórios entre os estudos podem decorrer dos diferentes procedimentos empregues para avaliar os aprendentes. De modo geral, são consideradas informações institucionais, estimativas baseadas no perfil linguístico e, mais raramente, são utilizados resultados de testes independentes. No presente estudo, foi desenvolvida uma versão portuguesa do *C-test*, um teste de preenchimento simples utilizado para medir a proficiência geral. Elevados coeficientes de *Cronbach Alfa* e *Omega* revelaram a fiabilidade do teste. A classificação da proficiência dos aprendentes baseada na pontuação do teste demonstrou uma correlação forte com a avaliação pelos professores ( $r = 0.74$ ), mostrando que o teste é eficiente para avaliar a proficiência geral. Para além disso, outros parâmetros, nomeadamente nível da turma e autoavaliação dos aprendentes também mostraram uma correlação forte com a avaliação por professores ( $r = 0.79$  e  $0.77$ ). A utilização futura do teste é discutida.

**Palavras-chave:** C-test. Avaliação. Proficiência. Aprendizagem de português. Português como língua estrangeira (PLE).

Proficiency assessment is crucial when one interprets the results of SLA studies. Inconsistent results among them can occur from a lack of uniformity in the methods of the proficiency assessment. In general, institutional status and estimates based on learners' linguistic profile are taken into consideration. However, the score of independent tests is rarely utilized. In this article, we developed a European-Portuguese version of the *C-test*, a simple fill-in-the-blank test used to measure general proficiency of a foreign language. High coefficients of Cronbach Alfa and Omega were found, demonstrating its reliability. Proficiency classification based on the test score showed strong correlation with the assessments made by teachers ( $r = 0.74$ ), which could imply that the test captures the general proficiency. Moreover, other parameters, namely classroom level and learners' self-assessment, also correlated strongly with the assessment by teachers ( $r = 0.79$  and  $0.77$ ). The use of the test for the future is discussed.

---

\* Universidade do Minho, Portugal.

**Keywords:** C-test. Assessment. Proficiency. Acquisition of Portuguese. Portuguese as a foreign language.



## 1. Introdução

Há várias décadas que um grande número de estudos (e.g. Anderson 1992; Bialystock & Smith 1985; Corder 1967; DeKeyser 1997; Ellis 2015; Firth & Wagner 2007; Isabel 2006; Krashen 1982; Lado 1957; Leiria 1991; Li 2010; Mcdonough & Kim 2009; Pienemann 1998; Pinto 2014; Rodrigues 2015; Selinker 1972; VanPatten 2007) tem sido realizado no que toca à aprendizagem de língua segunda e estrangeira<sup>1</sup> (doravante ALS). Através desses estudos foram revelados, por exemplo, vários fatores que afetam a variabilidade do desempenho e do nível de proficiência dos aprendentes, tais como língua materna, idade de início da aprendizagem da língua-alvo, duração da aprendizagem (em anos), frequência de uso da língua-alvo, entre outros. Estas variáveis têm sido tendencialmente aproveitadas para determinar a proficiência da L2. Sobretudo, o nível da turma e a duração da aprendizagem são as estimativas mais comuns na área da ALS, apesar da inexistência de homogeneidade em proficiência que os aprendentes apresentam nestes grupos (Tremblay 2011). Sendo assim, é necessário avaliar a proficiência de forma mais precisa e cuidadosa, por exemplo, através de um ou mais testes independentes.

Thomas (1994; 2006) examinou os métodos de aferição da proficiência de aprendentes em estudos publicados durante dois intervalos (1988–92 e 2000–04), em quatro revistas de ASL: 1) *Applied Linguistics*, 2) *Language Learning*, 3) *Second Language Research* e 4) *Studies in Second Language Acquisition*. O autor classificou esses métodos em quatro tipos: i) juízo impressionista (*impressionistic judgment*)<sup>2</sup>, ii) informações institucionais (*Institutional status*), iii) avaliação interna (*in-house assessment*) e iv) teste padronizado (*standardized test*). A percentagem de utilização das

---

1 Na área da ALS, o termo “língua segunda” refere-se, frequentemente a uma língua adicional: segunda, terceira, quarta e assim por diante (Ellis 2005). No presente trabalho, utilizam-se os termos língua segunda (L2) e língua estrangeira (LE) como sinónimos.

2 Trata-se de uma aferição subjetiva sem dados de suporte (p. ex., Os participantes deste estudo são principiantes.) ou com base em anos da estadia em locais em que se fala a língua-alvo.

*informações institucionais* (ii) foi de 40,1% no primeiro período e de 33,2% no segundo período; a de testes independentes (*avaliação interna e teste padronizado*) foi de 36,3% e 42,6%, respetivamente.

Também Trembley (2011), numa revisão de estudos sobre L2 publicados entre 2000 e 2008, indica que apenas 37,2% utilizaram um teste independente. Nos 62,8% dos estudos que não utilizaram testes independentes, 60,4% empregaram os parâmetros nível da turma e duração da aprendizagem para classificar o nível de proficiência dos participantes.

Deste modo, apesar de se verificar a utilização de testes independentes, este tipo de aferição ainda não pode ser considerado comum na área. Este cenário é mais saliente relativamente a algumas línguas. Por exemplo, Trembley (2011) relata ainda que, quanto aos estudos sobre o francês como L2/LE, o número é mais reduzido, somando apenas três. Algumas razões possíveis para a utilização reduzida de testes independentes são o facto de serem pagos e demorados e o número reduzido de instrumentos desenvolvidos para o efeito. Além disso, Lee-Ellis (2009) aponta o facto de que, ao contrário de línguas comumente investigadas como o inglês, não existe nenhuma medida “prática” de proficiência para o coreano, por exemplo, e, por conseguinte, testes independentes nos estudos de coreano como L2 são pouco utilizados.

Quanto ao português, o cenário é semelhante ao caso do francês e do coreano. Fizemos uma pesquisa de estudos sobre português em duas revistas, *Second Language Research* e *Studies in Second Language Acquisition*. Seguindo os critérios de Thomas (2006), foram excluídos: i) revisão de literatura ou livro, ii) ensaio, comentário ou outros tipos de estudo, em que se discute uma temática geral e não se realiza um estudo empírico com dados recolhidos. Adicionalmente, como o nosso interesse está voltado para a aprendizagem tardia, ainda foram excluídos estudos referentes à aprendizagem precoce (por crianças) e à de bilingues. Na *Second Language Research*<sup>3</sup> foram encontrados 85 resultados para a palavra “*Portuguese*”, sem determinar o período. Apenas o estudo de Montrul *et al.* (2010) foi enquadrado nos critérios de pesquisa estabelecidos.<sup>4</sup> Os autores avaliaram a proficiência de ingleses e espanhóis, aprendentes de português, a partir da autoavaliação e da duração da aprendizagem, ou seja, utilizando o critério do “juízo impressionista” de Thomas (2006). Já na *Studies in Second Language*

3 <http://journals.sagepub.com/home/slr> (Consultado em: 7 de março de 2019).

4 Dos 85 estudos apenas três continham a palavra em questão, isto é, “*Portuguese*”, no seu título. É provável que a palavra tenha sido encontrada no corpo do texto dos artigos.

*Acquisition*<sup>5</sup>, foram encontrados 16 resultados, sendo que nenhum se encaixava nos critérios indicados acima.<sup>6</sup> Consultamos, de seguida, a página da *Internet, Cátedra Português Língua Segunda e Estrangeira*<sup>7</sup>, que oferece uma lista de referências sobre aprendizagem e ensino de português L2. Dos 417 estudos listados, 277 foram descarregados a partir da ligação *download*.<sup>8</sup> Considerados os nossos critérios, restaram 55 estudos. A Tabela 1 mostra a utilização dos quatro métodos da classificação de Thomas (2006). As informações institucionais são maioritariamente utilizadas (67%), seguido pelo método juízo impressionista (25%), sendo utilizado algum teste independente apenas em 7% dos estudos.<sup>9</sup>

**Tabela 1. Utilização de quatro métodos de avaliação de proficiência**

<b>Tipo</b>	<b>Número</b>	<b>%</b>
juízo impressionista	14	25%
informações institucionais	37	67%
avaliação interna	3	5%
teste padronizado	1	2%
total	55	

Obviamente, a utilização de cada método pode ser legítima, dependendo do objetivo de investigação, pelo que não se pode afirmar que todos os estudos apresentem um problema metodológico. Porém, uma vez que i) há, frequentemente, dentro da turma, heterogeneidade em proficiência de

5 <https://www.cambridge.org/core/journals/studies-in-second-language-acquisition> (Consultado em: 7 de março de 2019)

6 Há alguns estudos que investigam a aprendizagem precoce ou por bilingues, ou que relatam uma temática geral sem recolher dados novos. Entre outros, por exemplo, Major (2007) investigou a identificação de acento de línguas familiares ou não familiares. Neste estudo, participaram americanos com/sem experiência do português. Os participantes com experiência podem ser considerados como aprendentes de português. No entanto, como não foi possível aceder aos textos integrais, apenas o resumo não apresentava informações suficientes para a inclusão na análise.

7 [http://catedraportugues.uem.mz/?\\_\\_target\\_\\_=bibli&bib=7](http://catedraportugues.uem.mz/?__target__=bibli&bib=7) (Consultado em: 7 de março de 2019)

8 Embora houvesse a ligação *download*, alguns artigos não foram obtidos por uma questão de direito de acesso.

9 Julgamos que Correia (2011) utilizou “teste padronizado”, visto que os textos analisados foram produzidos “num contexto de avaliação/certificação” do nível B2 do CAPLE, exame padronizado de português como língua estrangeira. No entanto, os textos foram produzidos pelos examinandos do exame DIPLE, equivalente ao nível B2, e não fica claro se todos possuem ou já obtiveram esse nível.

aprendentes e ii) cada instituição utiliza exames de posicionamento diferentes (isto é, existe também heterogeneidade em proficiência entre as turmas do mesmo nível), não se pode comparar nem generalizar os resultados dos estudos realizados, o que constitui um problema nos estudos da ASL.

Em português europeu, existe um teste padronizado, CAPLE<sup>10</sup>, efetuado nos Centros de Avaliação de Português como Língua Estrangeira para certificação da competência da língua. Este exame não é, no entanto, tal como os exames padronizados de outras línguas, um instrumento prático para estudos da ASL, já que é pago e demorado (normalmente demora-se um ou dois dias). Sendo assim, não temos instrumentos de acesso fácil para avaliar a proficiência de aprendentes para estudos da ASL, e pensamos que é vantajoso desenvolver um que possa ser utilizado de forma geral e económica no país. No presente trabalho, desenvolvemos uma versão em português europeu do *C-test*, um teste de preenchimento simples utilizado em várias línguas para medir a proficiência geral.

## 2. C-test

O *C-test* é um teste considerado como medida de proficiência geral de língua, que se baseia no princípio de redundância reduzida (Spolsky 1973), correspondendo, por exemplo, aos *noise test* e *cloze test* (Taylor 1953). De acordo com este princípio, considera-se que as mensagens linguísticas transmitidas no dia-a-dia contêm mais informações do que as que são rigorosamente necessárias. Por esta razão, mesmo que uma parte das informações se perca, consegue-se recuperar ou restituir a mensagem a partir das informações que estão intactas (Raatz & Klein-Braley 2002).

O *C-test* foi proposto por Raatz e Klein-Braley (1981) com a intenção de ultrapassar as limitações do *cloze test*. O teste consiste em 4 – 6 textos curtos autênticos. A primeira frase é mantida intacta e, a partir daí, são eliminados caracteres correspondendo a metade de cada palavra, de modo alternado (palavra sim palavra não) (Klein-Braley 1997). Quando o número de caracteres da palavra é ímpar, a maior parte dos caracteres é eliminada. Caso a palavra contenha apenas uma letra como “a”, “o”, “e”, é ignorada na contagem. A parte eliminada é substituída por um sublinhado de tamanho constante, isto é, independentemente do número de caracteres eliminados, conforme o exemplo infra (Baghaei & Tabatabaee 2015).

---

10 <http://caple.letras.ulisboa.pt>

*If you were to ask most people who Charles Darwin was, many of them would reply that he was the man who said that we were descended from monkeys. They wo\_\_\_ be wr\_\_\_. Darwin d\_\_\_ no mo\_\_\_ than sug\_\_\_ the possi\_\_\_. What h\_\_\_ said, a\_\_\_ proved b\_\_\_ thousands o\_\_\_ examples, w\_\_\_ that ov\_\_\_ millions o\_\_\_ years ani\_\_\_ and pla\_\_\_ have cha\_\_\_. This he called evolution.*  
(retirado de Baghaei & Tabatabaee 2015)

Os textos preparados são previamente verificados por falantes nativos da língua alvo. Consideram-se para utilização no teste apenas os textos cujas taxas de acerto sejam superiores a 90%. Os participantes devem inserir a parte eliminada, ou seja, restituir a palavra original. Apenas a restituição completa é considerada como resposta correta, pelo que, os erros ortográficos, por exemplo, podem ser tratados como resposta incorreta.

Desde a sua proposta, o teste foi traduzido para mais de 20 línguas (Eckes & Baghaei 2015), contando mais de 500 publicações (Grotjahn 2016). Apesar de haver controvérsia no que respeita à sua validade, muitos estudos apoiam a sua utilização do teste (p. ex., Babaii & Ansary 2001; Babaii & Moghaddam 2006; Eckes & Grotjahn 2006; Katona & Dornyei 1993; Klein-Braley 1997; Lei 2008). Outros autores, no entanto, consideram que o teste é demasiado fácil, revelando valores baixos de discriminação de itens (p. ex., Cleary, 1988; Kamimoto, 1993). Ainda outros julgam que o teste é adequado como medida da competência em nível micro (p.ex., leitura e gramática) mas não como medida da proficiência geral (p. ex., Chapelle & Abraham 1990; Cohen 1984).

De modo geral, a validação do *C-test* tem sido efetuada através da abordagem correlativa.<sup>11</sup> Muitos investigadores relatam correlação moderada ou forte entre o teste e outros tipos de testes considerados válidos. Além disso, apesar de o teste se correlacionar bem com vários componentes linguísticos a nível micro (p. ex., vocabulário, gramática, fala, escrita), apresenta melhor correlação com a pontuação total dos testes, isto é, a nível macro (Babaii & Ansary 2001; Eckes & Grotjahn 2006; Grotjahn & Stemmer 2002; Katona & Dornyei 1993), o que implica que o seu constructo do teste se baseia na avaliação integrativa.

11 Também se investiga através de abordagem fatorial (p. ex., Eckes & Grotjahn 2006; Khodadady 2014; Klein-Braley 1997).

### 3. Método

#### 3.1. Material

O teste foi criado seguindo os procedimentos propostos por Raatz & Klein-Braley (2002). Foram preparados cinco textos. Para que o teste abarcasse proficiências distintas, os textos foram retirados de manuais didáticos<sup>12</sup> e de jornais. A inteligibilidade<sup>13</sup> dos textos (Curto 2014) foi verificada com recurso à ferramenta LX-CEFR (Curto, Mamede & Baptista 2014), sendo que o valor de cada texto corresponde aos níveis A1, A2, B1, B2 e C1, respetivamente. Os textos são ordenados desde o mais fácil até ao mais difícil, para que os textos mais complexos não desmotivem os alunos de nível baixo logo no início do teste. Cada texto contém 20 lacunas. Cada lacuna foi criada, como habitualmente, pela regra mencionada acima: a partir da segunda frase, elimina-se a metade de palavras alternadas. Quando o número de caracteres da palavra é ímpar, a maior parte dos caracteres é eliminada. A parte eliminada é substituída por um sublinhado de tamanho idêntico. O texto de exemplo pode ser visto abaixo:

*O Manuel é estudante na Universidade de Lisboa. Ele lev\_\_\_\_\_ -se(1) sempre mu\_\_\_\_\_ (2) cedo. D\_\_\_\_\_ (3) manhã e \_\_\_\_\_ (4) tem au \_\_\_\_\_ (5) das 8h à \_\_\_\_\_ (6) 12h. Dep \_\_\_\_\_ (7) almoça n \_\_\_\_\_ (8) cantina c \_\_\_\_\_ (9) os col \_\_\_\_\_ (10). À ta \_\_\_\_\_ (11), o Manuel pra \_\_\_\_\_ (12) desporto: basqu \_\_\_\_\_ (13). Ele jo \_\_\_\_\_ (14) na equ \_\_\_\_\_ (15) da univer \_\_\_\_\_ (16). À no \_\_\_\_\_ (17), o Manuel ja \_\_\_\_\_ (18) com a fam \_\_\_\_\_ (19). Depois d \_\_\_\_\_ (20) jantar o Manuel gosta de navegar na internet ou falar com os amigos.*

O teste foi aplicado primeiro a 10 falantes nativos de português a fim de verificar se seria fácil para quem tem proficiência alta da língua. Todos os nativos conseguiram preencher corretamente quase todas as lacunas.<sup>14</sup>

12 Isto significa que nem todos os textos preparados eram autênticos. No entanto, o autor decidiu extrair os textos de manuais para assegurar que estivessem adequados aos níveis A1 e A2, que também se pretendiam considerar na avaliação do instrumento.

13 Equivalente ao termo inglês “*readability*” (Flesch Reading Ease; Flesch 1948).

14 O teste completo encontra-se em: <https://goo.gl/AVmTwv>.



### 3.2. Participantes

Participaram 104 aprendentes de português, que aprendem a língua em universidades no país. A maioria dos aprendentes eram alunos do curso de português para estrangeiros em universidades e alguns eram alunos de mestrado. A Tabela 2 apresenta o resumo do perfil linguístico de acordo com o nível da turma a que pertencem, isto é, uma informação institucional.<sup>15</sup>

Tabela 2. Perfil linguístico de acordo com a informação institucional

Nível da turma	N	Anos de aprendizagem	Média de idade (anos)	Nacionalidade <sup>16</sup>
B1	28	1,83	26,7	CN(10), MO(1), HK(1), VE(7), NL(1), AR(1), FR(2), SY(1), US(1), ES(1), IT(1), TZ(1)
B2	32	4,02	28,6	CN(11), US(4), CA(1), CH(1), JP(1), DE(2), KR(2), FR(1), UA(1), SY(2), EN(1), NL(1), RO(1), ES(1), RU(1), PT(1)
C1	38	6,34	23,3	CN(22), RU(3), US(1), BY(1), IE(1), ET(1), JP(1), DE(1), PT(3), BR(2), MZ(1), ES(1)
M <sup>17</sup>	6	5,88	34,0	CN(4), KR(1), JP(1)

### 3.3. Procedimento

O teste foi aplicado durante as aulas ou individualmente. Antes do teste, os participantes assinaram o termo de consentimento informado e preencheram

<sup>15</sup> Alguns participantes exibiam, em relação ao português, uma proficiência quase nativa. Como se previa que pudessem alcançar pontuações próximas das obtidas pelos falantes nativos, pensou-se na possibilidade de excluir os seus dados da análise. No entanto, desta vez, optou-se por incluí-los, a fim de verificar se o teste demonstra, de facto, um resultado como o esperado. Efetivamente, todos os participantes com este perfil tiveram pontuação acima de 80 (sendo 100 o número máximo de pontos possíveis).

<sup>16</sup> A sigla da nacionalidade é baseada na lista de códigos de países usados pela OTAN. O número à direita refere-se ao número de participantes.

<sup>17</sup> Refere-se aos alunos de mestrado. Estes alunos não frequentavam um curso de português e não era possível atribuir-lhes um nível da turma (B1, B2 e C1), pelo que foram classificados no grupo M.

uma ficha de informações destinada a caracterizar o perfil da amostra (*cf.* o teste completo em <https://goo.gl/AVmTwv>). Procedeu-se, de seguida, à explicação do teste e a uma sessão de treino. O teste foi realizado em 30 minutos na presença do investigador para garantir que os participantes não usassem o dicionário ou qualquer material auxiliar.

### 3.4. Pontuação

Foi dado um ponto para as respostas corretas e 0 para as incorretas. Uma única resposta era aceite, considerando-se como resposta incorreta erros ortográficos e outras palavras que poderiam funcionar gramatical e semanticamente.<sup>18</sup> Assim, a pontuação máxima de cada texto são 20 pontos e o total dos cinco textos 100 pontos.

## 4. Resultados

A análise foi efetuada com recurso à ferramenta R (v. R 3.3.3) e Microsoft Excel® da forma que se descreve em seguida. Em primeiro lugar, analisou-se a estatística descritiva do teste, a fim de capturar a tendência geral de cada texto e da pontuação de cada grupo. Em segundo lugar, o teste foi analisado em termos da fiabilidade e a discriminação de itens. Em terceiro lugar, foi efetuada uma análise de *cluster* e considerou-se a possibilidade de o teste medir a proficiência geral dos aprendentes, em comparação com a avaliação feita por professores.

### 4.1. Estatística descritiva

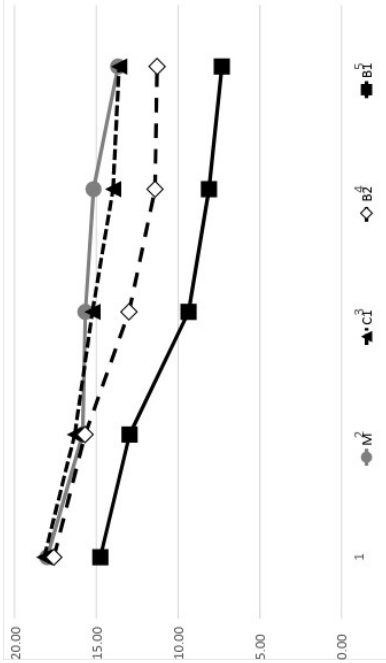
A estatística descritiva do teste é apresentada na Tabela 3. A tabela mostra que a média se vai tornando mais baixa do texto1 (T1) para o texto 5 (T5), o que implica que a dificuldade de cada texto varia conforme a sua inteligibilidade. Tal tendência é consistente entre os grupos, como é demonstrado no Gráfico 1.

---

<sup>18</sup> Em algumas questões foram permitidas respostas alternativas devido à variação de uso. Por exemplo, tanto *sande* como *sandes* foram considerados como corretos (2º texto, Nº 8). Além disso, não só *este* como também *esse* foram aceites (5º texto, Nº 7).

Tabela 3. Estatística descritiva do teste

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
T1	1	104	17.07	2.61	18	17.43	1.48	8	20	12	-1.20	0.98	0.26
T2	2	104	15.19	2.93	15	15.36	2.97	6	20	14	-0.61	0.45	0.29
T3	3	104	12.99	4.06	13	13.12	4.45	0	20	20	-0.35	-0.09	0.40
T4	4	104	11.67	4.59	12	11.86	4.45	0	20	20	-0.32	-0.53	0.45
T5	5	104	11.20	4.76	12	11.33	4.45	0	20	20	-0.28	-0.48	0.47



## 4.2. Fiabilidade

A fiabilidade é o critério que torna um teste consistente (Alderson 2005). É um parâmetro que garante que o teste pode apresentar o mesmo resultado independentemente da altura em que se aplica. Embora, tradicionalmente, sejam considerados vários métodos para estimar a fiabilidade, tais como o *test-retest*, *parallel forms*, *split-half* (para uma introdução, e.g. Hill & Hill 2008), o procedimento mais comum é o coeficiente *Cronbach Alpha* (Cronbach 1951). Em contrapartida, salienta-se que nos últimos anos alguns investigadores têm posto em causa o *Cronbach Alfa*, sendo que outro parâmetro tem sido recomendado (e.g. Okada 2011; 2015): o coeficiente *Omega*. Por estas razões, neste trabalho foram calculados ambos os parâmetros. A Tabela 4 mostra que o teste apresenta os coeficientes elevados. Acrescenta-se que os coeficientes foram calculados, considerando cada texto como um *super-item* ou *testlet* (Wainer & Kiely 1987), visto que se prevê a dependência local das lacunas num texto (Eckes & Grotjahn 2006; Klein-Braley 1985).

Tabela 4. Coeficientes de fiabilidade

<b>Alpha:</b>	0.93
<b>G.6:</b>	0.92
<b>Omega Hierarchical:</b>	0.87
<b>Omega H asymptotic:</b>	0.92
<b>Omega Total</b>	0.95

Para examinar se cada item tinha funcionado corretamente, foram calculadas a correlação item-total (*item-total correlation*, doravante CIT) e a dificuldade de item (doravante DI). A CIT representa o nível de discriminação, isto é, a correlação entre a pontuação de cada item e o total desse mesmo item, tendo o valor desde -1 até 1. O valor fica mais alto e aproxima-se de 1, caso o item tenha mais respostas corretas fornecidas por examinandos de pontuação alta e menos por aqueles de pontuação baixa. Um item cujo valor seja superior a 0,30 é considerado como tendo poder de discriminação (Brown 2005). A DI representa o quão difícil é o item, e calcula-se pelo “número de participantes com resposta correta / o número de participantes”. Quanto maior o valor, maior a facilidade do item. Os itens que se enquadram na categoria entre 0,30 e 0,70 são considerados de dificuldade intermédia.

A Tabela 5 apresenta os itens, cujo CIT se encontra abaixo de 0,30, juntamente com a sua DI à direita. A tabela mostra que a maioria dos itens é oriunda dos primeiros dois textos (Q1 – Q40). A baixa CIT destes itens leva-nos a presumir que os itens dos textos menos complexos, cuja inteligibilidade é A1 e A2 (T1 e T2), eram fáceis para todos os grupos (B1, B2, C1 e M). Visto que o teste foi criado para tentar capturar proficiências distintas de aprendentes, é lógico que haja tais itens (*e.g.* Q6, Q4, Q22). No entanto, alguns destes itens representam, ao mesmo tempo, a DI baixa (*e.g.* Q2, Q30, Q36), ou seja, maior dificuldade. Assim, pode-se afirmar que, mesmo nos textos de A1 e A2, havia alguns itens que eram bastante difíceis tanto para os aprendentes de nível intermédio como para os de nível avançado. Por esta razão, pode-se julgar que estes itens devam ser modificados. Em contrapartida, nenhum item demonstrou valor negativo. Por outras palavras, não havia itens problemáticos que fossem fáceis para os aprendentes de nível mais baixo, mas difíceis para os de nível mais avançado, o que nos leva a pensar que o teste foi, em termos gerais, bem construído.

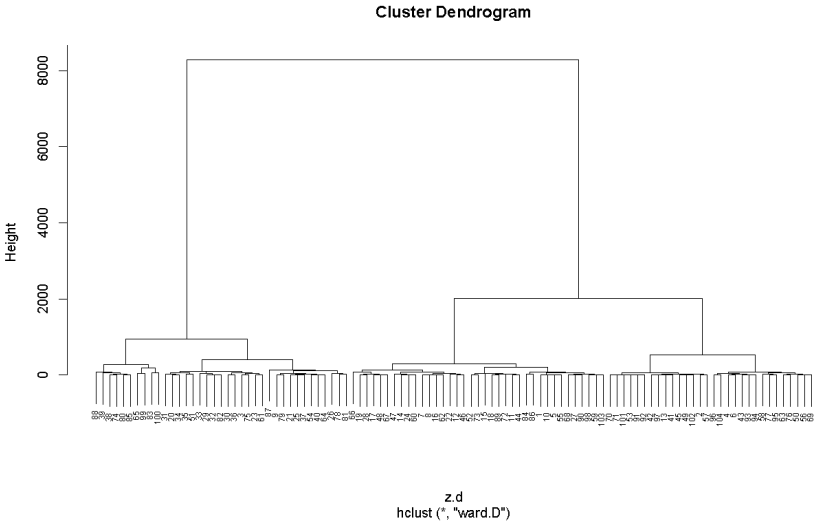
**Tabela 5. Itens de correlação item total baixa e a sua dificuldade**

<b>Item</b>	<b>CIT</b>	<b>DI</b>
Q2	0,12	0,23
Q30	0,14	0,28
Q36	0,15	0,28
Q29	0,16	0,31
Q15	0,19	0,35
Q13	0,19	0,32
Q53	0,20	0,37
Q17	0,21	0,31
Q40	0,22	0,36
Q73	0,23	0,29
Q61	0,23	0,30
Q83	0,23	0,33
Q28	0,23	0,29
Q6	0,24	0,85

Item	CIT	DI
Q27	0,24	0,30
Q16	0,25	0,42
Q25	0,25	0,36
Q35	0,26	0,38
Q26	0,27	0,45
Q4	0,27	0,54
Q80	0,27	0,40
Q22	0,28	0,62
Q5	0,29	0,49
Q11	0,29	0,66
Q9	0,29	0,77
Q20	0,29	0,62

### 4.3. Validade

Foi efetuada uma análise de *clusters* – método hierárquico – a fim de classificar a proficiência de acordo com a pontuação do teste. Foi obtido o dendrograma, apresentado no Gráfico 2. Tendo em conta a forma do dendrograma, foi decidido dividir os dados em quatro grupos, postulando os grupos como *principiante+*, *intermédio*, *intermédio+* e *avanzado*. De seguida, uma análise de *cluster* – método *k-means* – foi efetuada com 4 na variável *k*. A estatística descritiva de cada *cluster* encontra-se na Tabela 6, implicando que os *clusters* 1 a 4 representam *intermédio+*, *avanzado*, *intermédio* e *principiante+*, respetivamente. Para verificar se o agrupamento foi bem feito, uma ANOVA foi implementada em relação a cada variável (T1 – T5). A significância foi verificada entre todos os *clusters* (T1:  $F(3,100) = 51.097$ ,  $p < .001$ ; T2:  $F(3,100) = 31.842$ ,  $p < .001$ ; T3:  $F(3,100) = 112.653$ ,  $p < .001$ ; T4:  $F(3,100) = 114.968$ ,  $p < .001$ ; T5:  $F(3,100) = 105.776$ ,  $p < .001$ ).



**Gráfico 2. Dendrograma de cluster**

**Tabela 6. Estatística descritiva de cada cluster**

cluster 1		T1	T2	T3	T4	T5
	Min.	12.0	12.00	10.00	8.00	8.00
	1st Qu.	17.0	14.00	12.00	10.00	11.00
	Median	18.0	16.00	14.00	12.00	12.00
	Mean	17.7	15.49	13.76	11.73	12.27
	3rd Qu.	19.0	17.00	15.00	13.00	13.00
	Max.	20.0	20.00	17.00	15.00	16.00

cluster 2		T1	T2	T3	T4	T5
	Min.	17.00	15.0	12.0	14.00	12.00
	1st Qu.	19.00	16.0	16.0	16.00	14.00
	Median	19.00	17.5	17.5	17.00	16.00
	Mean	19.07	17.6	17.2	17.03	16.07
	3rd Qu.	20.00	19.0	19.0	18.00	18.00
	Max.	20.00	20.0	20.0	20.00	20.00

cluster 3		T1	T2	T3	T4	T5
	Min.	13.00	7.00	6.00	3.00	1.00
	1st Qu.	15.00	13.00	9.00	6.00	6.00
	Median	16.00	14.00	10.00	8.00	7.00
	Mean	15.81	13.78	9.852	8.30	7.15
	3rd Qu.	17.00	15.00	11.00	10.50	9.00
	Max.	19.00	18.00	12.00	13.00	11.00

cluster 4		T1	T2	T3	T4	T5
	Min.	8.0	6.0	0.00	0.00	0.00
	1st Qu.	11.0	10.0	4.50	1.75	0.75
	Median	11.0	11.0	6.50	5.50	4.50
	Mean	12.1	10.7	6.00	4.50	3.60
	3rd Qu.	12.0	12.0	7.75	6.75	5.75
	Max.	18.0	14.0	9.00	9.00	7.00

	cluster4	cluster3	cluster1	cluster2
Min. (total)	39,0	44,0	64,0	75,0
Max. (total)	43,0	65,0	80,0	97,0
Mean (total)	41,7	54,7	72,4	85,5

A Tabela 7 mostra a distribuição dos aprendentes em cada *cluster* conforme a sua informação institucional. Segundo estes dados, tudo indica que há bastante heterogeneidade dentro dos grupos. Por exemplo, nem todos os aprendentes do grupo C1 se enquadram no *cluster 2 (avançado)*, encontrando-se muitos no *cluster 1 (intermédio+)* também. A legitimidade deste agrupamento é discutida de seguida.



**Tabela 7. Distribuição dos aprendentes (cluster x grupo)**

Grupo	cluster1 (intermédio+)	cluster2 (avançado)	cluster3 (intermédio)	cluster4 (iniciante+)
B1	4	2	14	8
B2	13	8	9	2
C1	20	15	3	0
M	0	5	1	0

Recorde-se que a validação do *C-test* é feita principalmente através da abordagem correlativa, ou seja, a verificação de correlação com outros testes. Porém, em relação ao português, temos poucos instrumentos que poderiam servir de base de comparação. A maioria dos participantes não possuía o certificado do CAPLE, exame padronizado para certificação da competência de português, pelo que não havia nenhum parâmetro exterior que representasse a proficiência geral dos aprendentes. Assim sendo, no presente estudo, além da autoavaliação de aprendentes e do nível de turma (informação institucional), foi decidido utilizar a avaliação pelos professores. Os professores eram especialistas em PLE (português como língua estrangeira) e davam aulas aos aprendentes há algum tempo, pelo que as suas avaliações poderiam ser consideradas, até certo ponto, como um parâmetro confiável. A avaliação pelos professores foi obtida, considerando-se nove níveis, com base no QECR (Quadro Europeu Comum de Referência), mas de forma mais minuciosa para uma melhor captação da heterogeneidade de proficiência: A1, A1+, A2, A2+, B1, B1+, B2, B2+, C1, C1+, C2. O nível da turma corresponde na verdade a três níveis: B1, B2 e C1. A autoavaliação de alunos compreende 6 níveis: A1, A2, B1, B2, C1 e C2. A Tabela 8 apresenta a distribuição dos alunos conforme o nível obtido a partir da avaliação pelos professores. Pode-se considerar que há bastante heterogeneidade da proficiência dentro dos grupos. Se o *C-test* mede a proficiência geral, a sua pontuação também deve capturar tal heterogeneidade e correlacionar-se bem com a avaliação pelos professores. Os níveis de cada avaliação foram convertidos em escala numérica (ordinal) e foi calculada a correlação com a pontuação do teste.

**Tabela 8. Classificação da proficiência dos aprendentes com base na avaliação pelos professores**

Grupo <sup>19</sup>	A1	A1+	A2	A2+	B1	B1+	B2	B2+	C1	C1+	C2
B1	0	0	1	2	14	8	3	0	0	0	0
B2	0	0	0	0	3	5	10	8	5	0	1
C1	0	0	0	0	0	0	5	6	14	4	9

**Tabela 9. Correlação entre os quatro parâmetros examinados**

	C-test	Aval. Prof.	Autoaval.	N. Turma
C-test	1			
Aval. Prof.	0,74	1		
Autoaval.	0,71	0,79	1	
N. Turma	0,60	0,79	0,77	1

Como mostra a Tabela 9, a pontuação do *C-test* correlaciona-se melhor com a avaliação pelos professores do que com outros parâmetros, i.e., a autoavaliação e o nível da turma. Partindo do pressuposto de que a avaliação pelos professores é uma medida confiável de proficiência dos aprendentes, este valor de correlação suporta, até certo ponto, a validade do teste. Porém, note-se que, ao mesmo tempo, tanto a autoavaliação como o nível da turma apresentaram correlação forte com a avaliação pelos professores.

Para capturar mais detalhadamente a relação entre as avaliações, foi calculado o coeficiente de *Kappa*, que mede o grau de concordância de avaliações nominais. Para tal, devido à discrepância de escala, cada avaliação foi convertida em quatro escalas nominais. A Tabela 10, abaixo, apresenta o coeficiente de *Kappa* entre a avaliação pelos professores e outras avaliações, juntamente com a proporção de concordância simples. O coeficiente de *Kendall*, que mede a proporção de concordância de escalas ordinais, também foi calculado, convertendo cada avaliação em escala ordinal. Os resultados de ambos os testes mostraram que os três parâmetros apresentaram correlação moderada e forte com a avaliação pelos professores.

<sup>19</sup> Para os alunos de mestrado não foi possível obter a avaliação pelos professores, pelo que foram excluídos da análise.

Tabela 10. Coeficiente de *Kappa* e a proporção de concordância simples

	C-test	N. Turma	Autoavaliação
avaliação p/ prof.	0.408 ***	0.532 ***	0.459 ***
proporção de conc.	58.2%	68.4%	62.2%

Tabela 11. Coeficiente de concordância de Kendall

	C-test	N. Turma	Autoavaliação
avaliação p/ prof.	0.838 ***	0.897 ***	0.877 ***

## 5. Discussão

No que toca à fiabilidade do *C-test*, foram obtidos valores elevados dos coeficientes de *Cronbach Alpha* e *Omega*. Apenas com estes parâmetros, não se pode afirmar com segurança que o teste é confiável, mas eles sustentam a sua fiabilidade. Em relação à análise da discriminação de item, alguns itens foram considerados como não apropriados, daí ser melhor ter em consideração a sua modificação (cf. Jafarpur 1999). Em termos de dificuldade, não houve muitos itens demasiado difíceis. No entanto, como há aprendentes cuja pontuação dos últimos dois textos foi de 0 (cf. T4 e T5 na Tabela 6), é bem provável que tais itens estivessem a levantar bastantes dificuldades aos alunos de nível mais baixo, sobretudo os do grupo B1. Eventualmente, esta versão do *C-test* talvez não seja apropriada para os grupos de nível baixo e intermédio.

Quanto à sua validade, a interpretação é difícil. Aparentemente, o teste assinalou a heterogeneidade em proficiência que podia existir dentro dos grupos classificados de acordo com o nível da turma. Simultaneamente, a sua correlação com a avaliação dos professores é forte bem como os outros parâmetros. Este resultado implica que o teste funciona até certo ponto, mas, ao mesmo tempo, não garante a sua superioridade em relação aos outros parâmetros: autoavaliação e o nível da turma. Além disso, o valor de correlação do teste é ligeiramente mais baixo do que os valores dos demais parâmetros, o que nos leva a considerar algumas limitações possíveis.

Em primeiro lugar, é provável que o teste não esteja a medir a proficiência geral, mas sim alguns componentes micro e específicos, como, por exemplo, o conhecimento do vocabulário ou o conhecimento gramatical. A Tabela 12 apresenta a correlação da pontuação do *C-test* com a autoavaliação dos alunos e avaliação pelos professores em termos de quatro

competências linguísticas: ouvir, falar, ler e escrever. Apesar de não ter sido verificada uma tendência saliente, a modalidade de leitura, em particular, parece apresentar uma melhor correlação com o teste.

**Tabela 12. Correlação de quatro modalidades da autoavaliação e da avaliação pelos professores com o C-test**

	autoavaliação					aval. prof.				
	ouvir	falar	ler	escrever	total	ouvir	falar	ler	escrever	total
c-test	0,68	0,62	0,69	0,60	0,69	0,75	0,75	0,78	0,77	0,74

Em segundo lugar, é possível que o agrupamento pela análise de *cluster* não tenha sido bem feito. Como mostra a Tabela 6, o *cluster* 1, que é considerado como *intermédio+*, tem um traço diferente dos outros: a pontuação do T5 é ligeiramente melhor do que a do T4. Assim, alguns alunos que tiveram pontuação elevada (mais de 80 pontos) foram classificados no *cluster* 1 (*intermédio+*) e não no *cluster* 2 (*avançado*), provavelmente, por causa da distribuição das suas pontuações. Da mesma maneira, outros aprendentes com pontuação mais baixa do que a daqueles foram classificados no *cluster* 2 (*avançado*). De modo geral, na análise de *cluster* são usados vários métodos em termos da distância entre os dados e do agrupamento de *cluster*, e não há um método absolutamente fiável. No presente estudo, usamos o quadrado da distância Euclidiana e o método *Ward*. Talvez pudéssemos ter obtido outro resultado com outros métodos. Por outro lado, visto que a maioria dos aprendentes que tiveram mais de 80 pontos foi considerada pelos professores como estando no nível C1, C1+ ou C2, talvez seja melhor traçar linhas de referência para classificar a proficiência sem depender do agrupamento por *clusters*.

Por fim, deve-se considerar a questão do peso pontual. O grupo incluía muitos aprendentes classificados em níveis diferentes devido a uma diferença de apenas alguns pontos. Supondo que todos os que obtiveram mais de 80 pontos podiam ser considerados como estando no nível *avançado*, será que os que obtiveram 78 ou 79 correspondem realmente ao nível *intermédio+*? Provavelmente, o teste pode distinguir os aprendentes de nível avançado dos aprendentes de nível mais baixo, mas dificilmente poderá distinguir os aprendentes que estejam próximos da fronteira entre dois níveis. Recorde-se que, muito provavelmente, o teste foi bastante difícil para os aprendentes de nível mais baixo. Sendo assim, talvez seja mais apropriado que o teste seja utilizado para verificar se os aprendentes exibem ou não proficiência a um nível avançado.

## 6. Considerações finais

O presente estudo examinou a fiabilidade e a validade de uma versão em português europeu do C-test. Foi constatada a fiabilidade elevada através da Teoria de Teste Clássica. Por outro lado, é de salientar que, nos últimos vinte anos, os investigadores tentaram examinar o *C-test* através da Teoria de Resposta ao Item (TRI; *Item response theory*), mais concretamente através de *Polytomous Rasch Models* (Masters 1982; Samejima 1968) e *Testlet Response Theory* (Wang & Wilson 2005). Para tais análises, no entanto, são necessários muitos dados. Por esta razão, neste trabalho, que contém como amostra cerca de 100 participantes, não foi efetuada a análise do ponto de vista da TRI, o que é uma das limitações do estudo. Com a recolha de mais dados e uma análise baseada na TRI poderíamos obter mais informações em relação à fiabilidade e aos itens ou textos.

Em relação à validade, esta versão do teste ainda não pode ser considerada para avaliar a proficiência geral, visto que se implicou uma maior relação com a competência da leitura, bem como a impossibilidade de agrupamento minucioso. Sendo assim, chegamos à conclusão de que seria mais seguro utilizar a avaliação feita pelos professores de PLE. Naturalmente, esta nem sempre estará disponível, por exemplo, no caso de os aprendentes serem alunos de licenciatura ou mestrado, que não frequentam um curso de PLE. Nesses casos, pode-se utilizar o teste com algumas limitações como, por exemplo, apenas para verificar a proficiência elevada.

O *C-test* tem sido controverso desde a sua proposta. Muito embora existam vários estudos a suportar a sua validade, é importante não sobrestimar o seu alcance. Pelo seu formato, evidentemente, não se pode considerar um teste suficientemente robusto para capturar todo o conhecimento dos aprendentes. É provável que o teste funcione para um grupo de aprendentes, mas não para outro (Tremblay 2011). Será importante, por isso, interpretar os resultados e o potencial do teste com cautela, de vários pontos de vista e com várias amostras, inclusive, porque a validação é um processo constante. Por outro lado, também se pode afirmar que, quando devidamente validado, o *C-test* será um instrumento útil, uma vez que é relativamente fácil de desenvolver e é aplicável em tempo curto. Espera-se que sejam realizados mais estudos no que toca à avaliação de proficiência de aprendentes, que é indispensável para estudos empíricos em ALS.

## Referências

- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Londres & Nova Iorque: Continuum.
- Anderson, J. (1992). Automaticity and the ACT theory. *The American Journal of Psychology*, 105(2), 165–180.
- Babaii, E. & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System*, 29(2), 209–219.
- Babaii, E. & Moghaddam, M. J. (2006). On the interplay between test task difficulty and macro-level processing in the C-test. *System*, 34(4), 586–600. <https://doi.org/10.1016/j.system.2006.09.002>.
- Baghaei, P. & Tabatabaee, M. (2015). The C-Test: An integrative measure of crystallized intelligence. *Journal of Intelligence*, 3(2), 46–58. <https://doi.org/10.3390/jintelligence3020046>.
- Bialystock, E. & Smith, M. S. (1985). Interlanguage is not a state of mind: An evaluation of the construct for second-language acquisition. *Applied Linguistics*, 6(2), 101–117.
- Brown, J. (2005). *Testing in Language Programs*. McGraw-Hill Companies.
- Chapelle, C. A. & Abraham, R. (1990). Cloze method: what difference does it make? *Language Testing*, 7(2), 121–146.
- Cleary, C. (1988). The C-test in English. *RELC Journal*, 19(2), 26–37.
- Cohen, A. D., Segal, M., & Bar-Siman-To, R. (1984). The C-Test in Hebrew. *Language Testing*, 1(2), 221–225. <https://doi.org/10.1177/026553228400100206>.
- Corder, P. (1967). The significance of learner's errors. *International Review of Applied Linguistics*, 5(1), 161–170.
- Correia, R. D. Z. (2011). *Os erros no discurso escrito de Hispano-Falantes de nível B2* (Dissertação de Mestrado, Universidade de Lisboa).
- Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Curto, P. (2014). *Classificador de textos para o ensino de português como segunda língua* (Dissertação de Mestrado, Universidade de Lisboa).
- Curto, P., Mamede, N. & Baptista, J. (2014). Automatic readability classifier for European Portuguese. *System*, 5, 6.
- DeKeyser, R. (1997). Beyond explicit rule learning. *Studies in Second Language Acquisition*, 19(2), 195–222.
- Eckes, T. & Baghaei, P. (2015). Using Testlet Response Theory to Examine Local Dependence in C-Tests. *Applied Measurement in Education*, 28(2), 85–98. <https://doi.org/10.1080/08957347.2014.1002919>.
- Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23(3), 290–325. <https://doi.org/10.1191/0265532206lt330oa>

- Ellis, N. (2015). Implicit and explicit learning.pdf. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 3–23). Amsterdão: John Benjamins.
- Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition*, 27(2), 141–172.
- Firth, A. & Wagner, J. (2007). Second/Foreign Language Learning as a Social Accomplishment: Elaborations on a Reconceptualized SLA. *The Modern Language Journal*, 91(s1), 800–819. <https://doi.org/10.1111/j.1540-4781.2007.00670.x>.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.
- Grotjahn, R. (2016). *The electronic C-Test bibliography: Version 2015*. Disponível em: <[http://www.c-test.de/deutsch/ctest/pdf/C%20Test%20Bibliography/Grotjahn\\_Electronic\\_Ctest\\_Bibliography.pdf](http://www.c-test.de/deutsch/ctest/pdf/C%20Test%20Bibliography/Grotjahn_Electronic_Ctest_Bibliography.pdf)>.
- Grotjahn, R. & Stemmer, B. (2002). C-Tests and language processing.pdf. In J. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University language testing and the C-Test* (pp. 115–130). Bochum: AKS-Verlag.
- Hill, M. & Hill, A. (2008). *Investigação por questionário*. Edições Sílabo.
- Isabel, L. (2006). *Léxico, aquisição e leitura do português europeu língua não materna*. Lisboa, Portugal: Fundação Calouste Gulbenkian.
- Jafarpur, A. (1999). Can the C-test be improved with classical item analysis? *System*, 27(1), 79–89.
- Kamimoto, T. (1993). Tailoring the test to fit the students : Improvement of the C-test through classical item analysis. *Fukuoka Women's Junior College Studies*, 30(11), 47–61.
- Katona, L. & Dornyei, Z. (1993). The C-test. *FORUM*, 31(2), 35–38.
- Khodadady, E. (2014). Construct Validity of C-tests: A Factorial Approach. *Journal of Language Teaching and Research*, 5(6). <https://doi.org/10.4304/jltr.5.6.1353-1362>.
- Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing*, 2(1), 76–104.
- Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, 14(1), 47–84.
- Krashen, S. (1982). *Principles and Practice in Second Language Acquisition*. Pergamon Press: Longman.
- Lado, R. (1957). *Linguistics across cultures*. Estados Unidos da América: The University of Michigan Press.
- Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, 26(2), 245–274. <https://doi.org/10.1177/0265532208101007>.
- Lei, L. (2008). Validation of the C-Test amongst Chinese ESL Learners. *THE JOURNAL OF ASIA TEFL*, 5(2), 117–140.
- Leiria, I. (1991). *A aquisição por falantes de Português Europeu língua não materna dos aspectos verbais expressos pelos Pretéritos Perfeito e Imperfeito* (Dissertação de Mestrado, Universidade de Lisboa).

- Li, S. (2010). The Effectiveness of Corrective Feedback in SLA: A Meta-Analysis: Meta-Analysis of Corrective Feedback. *Language Learning*, 60(2), 309–365. <https://doi.org/10.1111/j.1467-9922.2010.00561.x>.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- Mcdonough, K. & Kim, Y. (2009). Syntactic Priming, Type Frequency, and EFL Learners' Production of *Wh*- Questions. *The Modern Language Journal*, 93(3), 386–398. <https://doi.org/10.1111/j.1540-4781.2009.00897.x>.
- Montrul, S., Dias, R. & Santos, H. (2010). Clitics and object expression in the L3 acquisition of Brazilian Portuguese. *Second Language Research*, 27(1), 21–58.
- Okada, K. (2011). Beyond Cronbach's alpha: a comparison of recent methods for estimating reliability. *The Japanese Journal for Research on Testing*, 7(1), 38–50.
- Okada, K. (2015). Reliability in psychology and psychological measurement, with focus on Cronbach's Alpha. *The Annual Report of Educational Psychology in Japan*, 54(1), 71–83.
- Pienemann, M. (1998). *Language processing and second language development: Processability Theory*. Amsterdão: John Benjamins.
- Pinto, J. (2014). A aquisição do género e da concordância de género em português língua terceira ou língua adicional. In P. Osório & F. Bertinetti (Eds.), *Teorias e Usos Linguísticos*. Lisboa: Lidel.
- Raatz, U. & Klein-Braley, C. (1981). The C-testa – modification of the cloze procedure. *Practice and Problems in Language Testing, University of Essex Department of Language and Linguistics Occasional Papers No. 26*.
- Raatz, U. & Klein-Braley, C. (2002). Introduction to language and to C-Tests. In James Coleman, R. Grotjahn, & U. Raatz (Eds.), *University Language Testing and the C-Test*. AKS-Verlag Bochum.
- Rodrigues, E. (2015). Concordância de número e gênero em estruturas predicativas no português brasileiro. *Linguística*, 11(1), 135-152.
- Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Report Series*, 1968(1), i-169.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1-4), 209-241.
- Spolsky, B. (1973). What does it mean to know a language; or how do you get somebody to perform his competence? In J. Oller & J. Richards (Eds.), *Focus on the learner* (pp. 164-176). Newbury House Pub.
- Taylor, W. (1953). Cloze procedure - a new tool for measuring readability. *Journalism Quarterly*, 30(4), 414-438.
- Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, 44(2), 307-336.



- Thomas, M. (2006). Research synthesis and historiography. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279-298). Amsterdam & Philadelphia: John Benjamins Publishing Company.
- Tremblay, A. (2011). Proficiency assessment standard in second language. *Studies in Second Language Acquisition*, 33(3), 339-372. <https://doi.org/10.1017/S0272263111000015>.
- VanPatten, B. (2007). Input processing in adult SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 115-135).
- Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3), 185-201.
- Wang, W.-C. & Wilson, M. (2005). The Rasch Testlet Model. *Applied Psychological Measurement*, 29(2), 126-149. <https://doi.org/10.1177/0146621604271053>.

[recebido em 27 de abril de 2018 e aceite para publicação em 20 de dezembro de 2018]

# APLICAÇÃO DE FERRAMENTAS PARA COLETA E ANÁLISE DE DADOS EM LINGUÍSTICA

## AN APPLICATION OF TOOLS FOR DATA COLLECTION AND ANALYSIS IN LINGUISTICS

Roberlei Alves Bertucci\*

bertucci@utfpr.edu.br

O uso de tecnologias influencia diferentes campos de saber e atividade humanos, inclusive a linguagem. Nesse sentido, as potencialidades verificadas nos ambientes digitais para circulação de dados linguísticos precisam ser exploradas. Para isso, este trabalho descreve três ferramentas digitais relacionadas à coleta e análise de dados: primeiro, o aplicativo *Netvizz*, integrado ao *Facebook*, que auxilia na montagem de *corpora* com dados dessa rede; segundo, o *software Tropes*, capaz de analisar textos a partir do processamento lexical, indicando elementos como o estilo e frequência das categorias lexicais; finalmente, o programa *Linguakit*, que seleciona palavras-chave, apresenta a frequência de palavras e realiza análise de sentimentos, entre outras tarefas. Para testar as ferramentas, selecionamos um conjunto de dados, retirado da página do *El País Brasil* por ocasião da prisão do ex-presidente Lula. Após a coleta de comentários, a aplicação no *Tropes* mostrou uma ocorrência alta de conectivos e modalizadores, além de itens lexicais referentes à situação (“universo de referência”). Já a análise no *Linguakit* apontou, além da alta frequência de termos específicos na situação, elementos típicos da Comunicação Mediada por Computador (como abreviações), bem como um sentimento mais negativo associado aos comentários.

**Palavras-chave:** Linguagem e tecnologia. Coleta de dados. Análise automática de textos. Rede social.

Technologies modify different human activities and knowledge fields, including language. Thus, some potentialities verified on digital environments for linguistic data interaction must be explored. In order to discuss how to do that, this paper describes three digital tools related to data collection and analysis: firstly, *Netvizz* App, on Facebook, which contributes to organize *corpora* by using data from this social network; secondly, *Tropes software*, which analyses texts from a lexical process, describing elements like text style and word frequency; thirdly, *Linguakit* program, which selects keyword, presents word frequency and does some sentiment analysis, among other tasks. To show how these tools work, we collect data from

---

\* Universidade Tecnológica Federal do Paraná, Brasil.

*El País Brasil* on Facebook, at the day which former president Lula has arrested. After this collection, *Tropes* analysis presented a higher frequency of both connectives and modal expressions, besides lexical items related to that situation. In turn, *Linguakit* analysis described current elements of Computer Mediated Communication (like abbreviations), besides a higher frequency of specific situation expressions, as well as a negative sentiment related to those comments.

**Keywords:** Language and technology. Data collection. Automatic text analysis. Social network.



## 1. Introdução

Se a reflexão sobre o papel da língua numa comunidade perpassa boa parte da nossa história, especialmente a partir da Era clássica, sua relação com a tecnologia é uma discussão relativamente recente. Autores como Auroux (2014), defendem que a escrita pode ser considerada como a primeira grande revolução tecnolinguística: é a técnica, repleta de reflexão sobre a sua formação e viabilidade (tecnologia), que coloca a língua como item fundamental do processo de conhecimento. Continuando, o autor refere como segunda revolução técnico-linguística a gramatização das línguas, cujo processo de reflexão sobre as possibilidades de organização e interpretação gerou produtos como gramáticas e dicionários. A gramatização assume um papel fundamental no processo de contato entre línguas, sobretudo na época das expansões de povos, a partir do século XV. Sem a escrita, no entanto, a gramatização não poderia ocorrer, pois apenas com o registro de uma língua se pode conhecê-la e estudá-la de modo científico.

Os estudos sobre as línguas, desenvolvidos de modo mais acentuado a partir do século XX, foram possibilitados por inúmeras tecnologias, para além do aparecimento da escrita e da gramatização massiva. A invenção do gravador, por exemplo, foi um passo decisivo no desenvolvimento de trabalhos em fonética, fonologia e variação (entre várias outras áreas), já que o registro da fala permitiu o estudo de uma língua com recurso à criação de bancos de dados. O computador, por sua vez, transformou-se tanto em uma ferramenta de (re) produção e/ou análise da linguagem, como em uma ferramenta de organização de dados (da fala e da escrita). É neste contexto que o presente trabalho pretende descrever o apoio computacional para coleta e análise de dados.

A possibilidade de uso das ferramentas que aqui trazemos tem relação direta com o Manifesto das Humanidades Digitais (2012): à medida que a sociedade se vai configurando no ambiente tecnológico e as opções recaem sobre o digital, mudam as condições de produção e divulgação dos conhecimentos podendo, sem dúvida, retirar-se benefícios de pesquisas com qualidade que contribuem para o enriquecimento do saber – tal como se verificou com o aparecimento da escrita. Nesse sentido, importa mostrar como muito dos saberes e das potencialidades que envolvem ambientes digitais podem estar na mão dos próprios usuários-pesquisadores: participando em uma rede social, um linguista pode coletar dados suscetíveis de fundamentar teorias sobre como a linguagem funciona numa determinada comunidade (inclusive numa comunidade ‘digital’). Trata-se de uma justificativa que sustenta a elaboração do presente trabalho, já que, no caso acadêmico, se pode partir do princípio de que há uma relação bem estabelecida entre os estudantes e as tecnologias digitais, utilizadas para atividades cotidianas por meio da linguagem. Afinal, como afirma Coscarelli (2016, p. 11),

as tecnologias digitais, disponíveis agora nos celulares e amplamente utilizadas por todas as camadas sociais como meio de comunicação, produção e disseminação de saberes, precisam ser estudadas e compreendidas. Os mais diversos contextos escolares precisam discutir e se apropriar dessas tecnologias para que os alunos também incorporem em suas vidas as inúmeras possibilidades oferecidas por equipamentos e aplicativos.

Deste modo, esperamos também contribuir de forma efetiva para a discussão, apropriação e, quiçá, incorporação da relação entre linguagem e tecnologia. Aliás, na área de estudos de *corpora* tem-se destacado que, mais do que apenas processar mais rapidamente informações, o computador tem sido um valioso aliado dos estudos linguísticos, tanto do ponto de vista de coleta e análise de dados, quanto do ponto de vista de desenvolvimento de ferramentas vinculadas à linguagem.

Note-se que recursos de análise da língua portuguesa com recurso a ferramentas gratuitas não é uma novidade (Sardinha 2005). No entanto, o presente artigo procura mostrar que há a possibilidade tanto de coleta de dados em ambiente de redes sociais, como a possibilidade de processamentos desses *corpora* com ferramentas específicas, o que constitui uma novidade.

Por outro lado, é importante destacar que este trabalho não se insere necessariamente no âmbito de novas teorias para Linguística de Corpus, tendo uma intenção relativamente modesta: descrever como é possível formar *corpora* e analisá-los em ambientes computacionais. Assim, não

pretendemos discutir como um dado corpus *deve* ser formado, que critérios estão subjacentes à sua formação, mas sim, como ele *pode* ser construído nesses ambientes. Na descrição que se pretende fazer das ferramentas, algumas perguntas foram cruciais para traçar o plano de trabalho: 1) que possibilidades oferecem as ferramentas no que respeita ao estudo da linguagem? 2) como podem essas ferramentas ser utilizadas por pesquisadores interessados em dados de rede social ou em análises computacionais de textos?

Para encontrar as respostas, descrevemos uma coleta de dados realizada na rede social Facebook, por meio do aplicativo *Netvizz* nela integrado. Em seguida, apresentamos a análise desse corpus por meio dos *softwares* Linguakit e Tropes. Como se verá adiante, quer pelos resultados da coleta quer pelas análises, há uma grande facilidade de acesso por parte do pesquisador da linguagem que deseja tomar o ambiente virtual como espaço para coleta e pesquisa com finalidades linguísticas.

O presente artigo está dividido da seguinte forma: na Seção 1, analisamos a relação entre linguagem e tecnologia de forma mais aprofundada, apresentando a importância dos *corpora* para as análises em linguística; na Seção 2, descrevemos as ferramentas utilizadas, bem como as análises realizadas por meio delas; em seguida, apresentamos as considerações finais.

## 2. Linguagem, tecnologia e corpus

### 2.1. Da relação entre tecnologia e linguagem

Toda sociedade se modifica naturalmente, dadas as condições de desenvolvimento e estruturação que a envolvem e, por isso, é inegável que o conhecimento tecnológico modifica o modo de agir e de pensar das culturas: à medida que equipamentos são desenvolvidos (veja-se o caso do microscópio,), novos objetos de estudo passam a fazer parte do cotidiano científico daquele povo (o estudo de micro-organismos, por exemplo). Por isso, Cupani (2016) revela que a tecnologia pode, sim, influenciar o modo de pensar e os resultados das pesquisas em diferentes ambientes. Sem dúvida, a tecnologia é um tema e uma área que envolve a vida de todas as sociedades pois, como afirma Cupani (2016, p. 9), “a tecnologia nos afeta e desafia qualquer que seja nossa atividade”.

Por outro lado, nesse ambiente de afetação, a própria tecnologia passa a ser objeto de investigação: quando o ser humano encara a vida como um problema torna-se também capaz de procurar alternativas para o resolver (Ortega y Gasset 1939). Por isso, a inter-relação entre linguagem e tecnologia pode trazer perguntas (problemas) como: por que e para quê servem os

diferentes dispositivos criados? Ou ainda antes disso: o que permite que o homem desenvolva tais artefatos?

Uma possível resposta para esta última pergunta parece ser a relação entre tecnologia e planificação: como uma criação tem um objetivo, o artefato precisa de ser refletido. Por isso, a noção de planificação ganha destaque nessa perspectiva. De modo similar, Vieira Pinto (2005) argumenta que o planejamento é inerente à tecnologia; por outro lado, pode-se acrescentar o fato de que é a linguagem a facilitadora desse simbolismo (abstração) específico, capaz de permitir o planejamento e, conseqüentemente, a realização do artefato. Tal capacidade é, para Cassirer (1979, p. 49) a chave para se entender o próprio ser humano:

Entre o sistema receptor e o sistema de reação, que se encontram em todas as espécies animais, encontramos no homem um terceiro elo, que podemos descrever como o *sistema simbólico*; esta nova aquisição transforma toda a vida humana. Em confronto com os outros animais, o homem não vive apenas numa realidade mais vasta; vive, por assim dizer, numa nova dimensão da realidade. (grifos do autor)

Como se lê acima, o sistema simbólico, ofereceu ao homem condições claras de diferenciação dos outros animais, mas sobretudo, de superar dimensões mais próximas e desejar, planejar e até mesmo realizar outras, trazendo-as à dimensão real. Nesse sentido, portanto, a linguagem exerceu um papel decisivo, já que assume o papel de modeladora dessa capacidade.

Pode-se dizer que é graças a essa capacidade que as inovações modificam o ambiente humano. Barton e Lee (2015) consideram que toda mudança tecnológica causa alguma mudança na vida social; ou seja, à medida que novas técnicas, processos e produtos aparecem, a vida das pessoas se modifica. Como a linguagem é parte essencial no processo, não podemos desconsiderar a relação plena entre texto/linguagem e tecnologias: linguagem modificando práticas sociais (e tecnológicas) e tecnologias modificando/influenciando práticas linguísticas.

Portanto, para esses autores, “o mundo está cada vez mais mediado pelo texto e a web é parte essencial dessa mediação” (Barton & Lee 2015, p. 29). Sendo assim, não apenas se constata um espaço de circulação textual mais abrangente e mais ubíquo, como também se verifica a necessidade de um estudo mais específico sobre a mobilização dos recursos linguísticos utilizados nas práticas linguísticas nesse ambiente, o que é, sem dúvida, um desafio para os estudos sobre linguagens.

Igualmente importante é o fato de esse ambiente, que criou gêneros textuais relativamente instáveis ou rapidamente mutáveis (como os memes ou menes), ter permitido uma concentração de dados capaz de proporcionar

aos pesquisadores da linguagem uma boa fonte de compilação para suas pesquisas linguísticas. Dessa forma, ao mesmo tempo em que a Web fornece espaço para a criação textual, novos dispositivos de compilação de dados para sua análise são desenvolvidos em diferentes instituições.

Deste modo, é possível falar sempre de uma valoração da tecnologia, seus processos e produtos. Consequentemente, Cupani (2016, p. 12) argumenta que “aquilo que denominamos tecnologia se apresenta, pois, como uma realidade polifacetada: não apenas em forma de objetos e conjuntos de objetos, mas também como sistemas, como processos, como modo de proceder, como uma certa localidade”. Em outras palavras, se o mundo está cada vez mais permeado de tecnologia e a linguagem é a grande mediadora da sociedade, perante esta realidade lança-se o desafio aos estudiosos da área de identificarem objetos, processos e modos proceder na coleta e análise de dados.

Por conseguinte, procuramos descrever algumas possibilidades de utilização de ferramentas de coleta e análise textual, levando em conta tanto ambientes de grande circulação, como as redes sociais, como outros mais restritos, como os de atividades didáticas em sala. Mais especificamente, nossa intenção é descrever como se podem coletar dados de uma rede social por um aplicativo específico (*Netvizz*), bem como analisar tais dados a partir de outros aplicativos (*Linguakit* e *Tropes*).

Inegavelmente, a área de Linguística de Corpus tem se beneficiado do desenvolvimento de ferramentas capazes de gravar, armazenar e analisar dados de línguas naturais. A disponibilização dos dados eletrônicos bem como de ferramentas que permitem seu tratamento é um ponto fundamental para os resultados atingidos por pesquisadores da área. Como afirmam Raso e Melo (2012, p. 33):

A maior parte dos corpora produzidos no Brasil são escritos, principalmente com material pertinente a jornais e gêneros acadêmico-científicos, utilizados, sobretudo, por grupos de pesquisa voltados para os estudos do léxico e desenvolvimento de ferramentas computacionais para o tratamento da linguagem natural.

Das palavras dos autores depreende-se, portanto, que os corpora orais representam o português brasileiro em menor número, como também os corpora de textos menos monitorados são igualmente raros. Obviamente, aspectos operacionais para a construção e manutenção de banco de dados são fatores decisivos dessa realidade. No entanto, pode-se também falar de uma falta de interesse por “gêneros escritos menores”, na medida em que estes se situam no limiar entre a fala (a língua natural) e a escrita (a língua a ser

aprendida). Aqui, queremos assumir que o estudo de ambientes de escrita menos monitorados pode revelar-se muito produtivo para todas as áreas de estudos que tomam a linguagem (as relações sociais, portanto) como relevantes. Há, compreendemos, aspectos linguísticos bastante reveladores, não só no uso da estrutura da língua, como em sua força expressiva (ou discursiva).

Ainda que possa haver projetos a respeito de textos escritos em situações menos formais (textos de alunos de escola básica, por exemplo), pretendemos aqui destacar a possibilidade de estudos relativos a *corpora* com propósitos específicos, como é o caso dos textos produzidos em redes sociais, especialmente os comentários. Diferentes trabalhos têm dado o seu contributo com análises relevantes de fatores linguísticos em ambientes de interação *on-line* (Araújo & Leffa 2016; Barton & Lee 2015; Coscarelli 2016, entre outros), ainda que mais direcionados para as possibilidades pedagógicas.

Acreditamos, tal como ocorre com outros estudos que recorrem a *corpora*, que a utilização de textos produzidos em ambientes de interação digital pode fornecer elementos importantes para o estudo da língua, como seja a análise do discurso, o estudo de texto/gêneros, a variação linguística, entre outros. Paiva e Paredes-Silva (2012), por exemplo, discutem aspectos de variação e mudança observáveis na escrita, a partir de dados de textos jornalísticos. Ora, considerando que esse ambiente é bastante monitorado, poderíamos questionar se as descrições de variação aí observadas serão também encontradas em gêneros menos formais. Vale a pena referir, que a compilação de dados permanentes de tais gêneros não parece necessária, uma vez que não exige uma logística complexa, como no caso de gravações de textos orais. Por isso, sugerimos o uso de aplicativos, como o *Netvizz*, como grandes facilitadores do processo de compilação.

Importa salientar o uso de ferramentas computacionais em ambientes de ensino e/ou pesquisa, especialmente de materiais em língua. Finatto (2017) realça que, inegavelmente, houve uma série de avanços na compilação e análise de dados com apoio computacional em português nos últimos anos. Outro ponto fundamental é a importância das relações interdisciplinares geradas (ou exigidas) por tais contextos, já que especialistas dos sistemas de comunicação/computação estão em diálogo com linguistas.

Igualmente, Finatto (2017) destaca a importância do apoio computacional quer para a formação de *corpora* de diferentes modos e fontes, quer para sua análise textual, relevando a descrição de gêneros textuais/discursivos. Concordamos com a autora, pois tal apoio, especialmente para os linguistas, mostra que a tecnologia é um meio que ajuda a explicar diferentes fenômenos da língua natural.



O presente trabalho pretende contribuir para destacar o apoio dado por recursos tecnológicos na área da linguagem: a partir da coleta de dados em rede social, com o aplicativo *Netvizz*, procederemos à análise do *corpus* por meio de dois recursos: o *Linguakit* e o *Tropes*. Queremos descrever como um pesquisador pode encontrar nessas ferramentas um aliado na análise de *corpora* mais extensos.<sup>1</sup> A partir dos dados selecionados, indicaremos alguns recursos linguísticos recorrentes apontados pelos analisadores automáticos.

## 2.2. Linguística de Corpus

Apesar do presente trabalho, como dito, não fazer parte propriamente das pesquisas em Linguística de *Corpus*, o fato de descrever um modo de coleta de dados para análises linguísticas faz com que discutamos a relevância da montagem de corpora para pesquisas na área.

Em geral, a grande contribuição das pesquisas com *corpora* é permitir que os pesquisadores façam uma análise empírica, ou seja, de dados reais e, para isso, o uso de instrumentos de coleta e armazenamento de dados é essencial. Sardinha (2000, p. 325) destaca a importância dos recursos computacionais para a formação de *corpora* linguísticos.

A Linguística de Corpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador.

Como se vê, a própria concepção dos *corpora* está permeada de questões teóricas, já que o critério de formação do banco de dados é algo essencial tanto para a estrutura do *corpus* quanto para o acesso aos dados. Por isso, o recurso a meios computacionais é imprescindível. Na relação entre linguagem e tecnologia, o corpus é, portanto, um produto artificial (tecnológico), composto de uma série de etapas de reflexões e técnicas, cuja finalidade é contribuir para a pesquisa com dados empíricos sobre as línguas naturais.

---

1 Alguém poderia questionar se a limitação de caracteres de processamento de alguns *softwares*, como o *Tropes* e o *Linguakit*, não impediria análises mais extensas. Argumentamos que essa limitação não é tão grande: o *Tropes* processa até 32 mil caracteres e o *Linguakit* até 20 mil, o que é um número considerável, especialmente quando se trata de comentários on-line. Agradecemos a um parecerista que chamou a atenção para esse fato.

Dessa forma, para que seja possível a análise, é preciso que se estabeleçam alguns critérios. Sardinha (2000, p. 340–341) apresenta alguns dos principais pontos a serem levados em conta na concepção de um corpus, elencados e explicados no Quadro 1, a seguir.

CRITÉRIO	TIPO	COMPOSIÇÃO
<b>Modo</b>	Falado	porções de fala transcritas.
	Escrito	textos escritos, impressos ou não
<b>Tempo</b>	Sincrônico	compreende um período específico
	Diacrônico	compreende vários períodos
	Contemporâneo	representa o período corrente
	Histórico	representa um período do passado
<b>Seleção</b>	Amostragem ( <i>sample corpus</i> )	porções de textos ou de variedades textuais, planejado para ser uma amostra finita da linguagem como um todo
	Monitor	reciclada para refletir o estado atual de uma língua; opõe-se a corpora de amostragem
	Dinâmico ou orgânico	crescimento e diminuição são permitidos; qualifica o corpus monitor
	Estático	oposto de dinâmico; caracteriza o corpus de amostragem
	Equilibrado ( <i>balanced</i> )	os componentes (gêneros, textos, etc.) são distribuídos em quantidades semelhantes (por exemplo, mesmo número de textos por gênero)
<b>Conteúdo</b>	Especializado	tipos específicos (em geral gêneros ou registros definidos)
	Regional ou dialetal	uma ou mais variedades sociolingüísticas específicas
	Multilíngue	Inclui idiomas diferentes
<b>Autoria</b>	Aprendiz	Os autores dos textos não são falantes nativos
	Língua nativa	Os autores são falantes nativos
<b>Disposição interna</b>	Paralelo	Os textos são comparáveis (p.ex. original e tradução)
	Alinhado	As traduções aparecem abaixo de cada linha do original
<b>Finalidade</b>	Estudo	O corpus que se pretende descrever
	Referência	Usado para fins de contraste com o corpus de estudo
	Treinamento ou teste	Construído para permitir o desenvolvimento de aplicações e ferramentas de análise

**Quadro 1. Critérios de formação de *corpus*. Adaptado de Sardinha (2000, pp. 340-341)**

Na Seção 2, em que apresentamos a formação e análise do corpus, apontamos os itens específicos referidos no Quadro 2 que melhor se enquadram na perspectiva.

Do ponto de vista da representatividade, Sardinha (2000) alega que um *corpus*, independe do tamanho, mas sim que, tais conjuntos de dados, agrupados conforme critérios do pesquisador, precisam ser representativos do uso linguístico em alguma circunstância. Assim, não se pode definir uma extensão mínima para de um *corpus*, mas é essencial que seja entendido como suficientemente representativo. Além disso, embora não tenha extensão pré-definida, a representatividade é, logicamente, melhor caracterizada quanto maior for o *corpus*. Quando específicos (e não abertos), os *corpora* são de acesso exclusivo do pesquisador, que o produz para uma finalidade específica, não sendo disponíveis para outros pesquisadores e, conseqüentemente, acabam não sendo “verificáveis, o que compromete a pesquisa em termos de sua replicabilidade e generabilidade” (Sardinha 2000, p. 348).

No entanto, vamos mostrar, neste trabalho, que o *Netvizz*, sendo um aplicativo que coleta dados a partir da rede social pela conta do próprio usuário, torna essa especificidade não mais um obstáculo. Em outras palavras, um mesmo *corpus* pode ser usado por diferentes pesquisadores, ainda que não estejam armazenados num único local, nem mesmo sejam coletados no mesmo dia. Se os critérios forem idênticos, o *corpus* ficará acessível de forma ubíqua.

Na próxima seção, descrevemos com mais detalhes os recursos computacionais utilizados na pesquisa, bem como os resultados que fornecem a partir da compilação dos dados retirados do *Facebook*.

### 3. Ferramentas tecnológicas: compilação e análise de dados

Nesta seção, apresentamos algumas ferramentas computacionais que consideramos relevantes para o trabalho com *corpora*, especialmente aqueles que provêm das próprias redes sociais. Como apontado anteriormente, pretendemos verificar as potencialidades das ferramentas no que toca ao trabalho de pesquisa linguística, tanto do ponto de vista da coleta, quanto do ponto de vista da análise.

### 3.1. *Netvizz*: coleta de dados no Facebook

Para Araújo e Leffa (2016), o uso das redes sociais ou sua aplicação no ensino já vem sendo objeto de investigações graças a recursos que os promovem e à área que envolvem. Do ponto de vista das pesquisas sobre a língua, no entanto, aparece, ainda, haver carência de análises de dados relativos às suas variações/mudanças ou mesmo de aspectos já verificados em análises de outros textos orais e escritos.

Isso justifica o presente artigo, já que a forma como ocorrem as práticas de linguagem nesse ambiente é diferente da conversa presencial. Nesse sentido, se, como afirma Recuero (2012, p. 28), “num diálogo, tudo é informação: elementos prosódicos (como o tom da voz, a entonação e as pausas da fala), elementos gestuais e, evidentemente, as palavras”, no ambiente virtual todos os demais itens semiotizados precisam de ser interpretados, para que o leitor construa os sentidos ali presentes. Por isso, a navegação e a interação no meio digital, pela complexidade da convergência de diferentes semioses, distinguem-se da interação exclusivamente escrita. Essas questões abrem espaço inúmeras pesquisas, que podem envolver as reações numa publicação no Facebook, a forma de pontuação que ocorre em textos ali escritos, a construção de identidades no espaço virtual, ou mesmo a utilização de uma ferramenta de apreciação de um comentário, tudo em busca de compreender melhor como se dão as práticas e trocas sociais ali estabelecidas por meio da linguagem (Bertucci & Nunes 2017).

Embora o trabalho com redes sociais seja algo promissor nos estudos de linguagens, obter os dados dos materiais ali presentes, como postagens, reações, e comentários é algo bastante desafiador. Daí a relevância da apresentação do aplicativo *Netvizz* como meio facilitador dessa tarefa. Segundo a pesquisa realizada em abril de 2018 no Portal de Periódicos da Capes, o que melhor concentra trabalhos acadêmicos no Brasil, encontraram-se apenas três trabalhos que faziam referência ao aplicativo, dos quais apenas um na área de Linguística, que traçava uma relação próxima entre reações de raiva e divergência de opinião na rede (Bertucci & Nunes 2017). O desconhecimento do dispositivo justifica o presente trabalho, já que pode ajudar outros pesquisadores a coletarem dados da rede.

O *Netvizz* é um aplicativo integrado na rede social *Facebook*, disponibilizado a todos os usuários. Sua função é extrair dados de páginas e grupos que podem servir para pesquisas em diferentes áreas. Depois de extraídos, os dados precisam de ser consolidados em uma planilha específica, ou até sujeitos a análise por ferramentas de análise linguística. Embora esteja para

além do foco do nosso trabalho mostrar o passo a passo para sua utilização, apontamos os elementos necessários para a constituição de um *corpus* por meio desse aplicativo, sendo eles:

- a seleção da página fonte de dados;
- a seleção do período ou da quantidade de publicações da página;
- a consolidação em planilha (*e.g.* Excel);
- e a seleção/filtragem dos tipos de dados para análise.<sup>2</sup>

Para a compilação dos dados, o usuário precisa, em primeiro lugar, selecionar uma página ou grupo público de onde deseja extrair os dados. No nosso caso, extraímos dados da página do *El País Brasil*<sup>3</sup>, um periódico de notícias bastante atuante na rede. Dali, entendemos selecionar uma publicação de alto engajamento e comentários, por sua vez selecionando alguns deles para análise em outras ferramentas. Na sequência, acessamos o aplicativo *Netvizz*.

Depois, selecionamos o período: optamos pelos dias 7 e 8 de abril de 2018, data em que ocorreu a prisão do ex-presidente Lula da Silva. Consideramos que o clima de apreensão e tensão (e acirramento dos ânimos políticos) que envolveu o Brasil naquele período propício à produção de comentários com características linguísticas menos monitoradas. Embora nossa proposta de análise não seja especificamente de variação linguística, optamos por valorizar produções mais espontâneas, o que ocorre mais facilmente em situações de envolvimento emocional.

Em seguida importamos os dados gerados pelo aplicativo em uma planilha Excel, o que nos permitiu filtrar e selecionar a publicação com maior número de comentários (1.150). A Figura 1, que se segue, mostra um resumo da tabela consolidada gerada a partir do *Netvizz*, com as dez publicações mais comentadas na página nos dias 7 e 8 de abril (de um total de 31 publicações). Como se pode ver, os dados revelam um alto engajamento nas publicações (da menor, com 872, à maior, com 10.211); além disso, há um grande envolvimento dos usuários nessas postagens por meio de comentários (de 187 a 1.150).

---

2 Para uma descrição detalhada do aplicativo *Netvizz*, sugerimos o trabalho de Rieder (2013).

3 Disponível em: [www.facebook.com/elpaisbrasil](http://www.facebook.com/elpaisbrasil). Consultado em: 26 abr. 2018.

E		M	N	O	P	Q
		likes_count_fb	comments_count_fb	reactions_count_fb	shares_count_fb	engagement_fb
1	post_message	1979	1150	3381	1217	5748
2	Após a mobilização da militância diante da iminente prisão do ex-presidente o PT ainda não					
3	O Lula que volta à prisão 38 anos depois ainda conserva muitas coisas do ousado sindicalist	1731	504	2147	450	3101
4	Há apenas uma década tudo era muito diferente. Em 2008 enquanto a Europa e os EUA mer	4219	369	5864	2040	8273
5	A segregação espacial portua o dia a dia do Condomínio Laranjeiras onde um exército de se	1756	308	2991	1302	4601
6	O líder que protagonizou três décadas de política brasileira dedicou boa parte do discurso a	1612	280	2174	164	2618
7	Editorial   O Brasil deve realizar eleições num clima de estabilidade. Políticos juizes e milita	343	269	656	42	967
8	Segundo os dados coletados pelo grupo de Piketty a fatia do 1% mais rico de brasileiros fica	3786	241	5298	4672	10211
9	Repúdio ou silêncio. A prisão do ex-presidente Luiz Inácio Lula da Silva que se entregou à pr	1904	196	2255	344	2795
10	Ex-presidente fez um discurso por quase uma hora a apoiadores diante Sindicato dos Metalu	2261	188	2811	335	3334
11	Tudo começou em março de 2014 quando em uma operação rotineira sobre crimes financeir	419	187	550	135	872

Figura 1. Tabela consolidada – dados do Netvizz/Facebook

Fonte: O autor

Isso feito, decidimos submeter uma parte dos comentários da postagem selecionada (149) para análise nas ferramentas *Tropes* e *Linguakit*. Nossa intenção era verificar quais elementos linguísticos poderiam ser destacados desse *corpus* com 31 publicações.

Antes de passarmos à análise, propriamente dita, poderíamos classificar o *corpus* que montamos a partir da tipologia apresentada.

CRITÉRIO	TIPO	NETVIZZ/CORPUS ATUAL	EXPLICAÇÃO
<b>Modo</b>	Falado	escrito	Textos escritos (comentários) de rede social.
	Escrito		
<b>Tempo</b>	Sincrônico	contemporâneo	Representa o período corrente (07-08/04/2018)
	Diacrônico		
	Contemporâneo		
	Histórico		
<b>Seleção</b>	Amostragem ( <i>sample corpus</i> )	amostragem	Com porções de textos ou de variedades textuais do Facebook, foi planejado para ser uma amostra finita da linguagem como um todo naquele ambiente.
	Monitor		
	Dinâmico ou orgânico		
	Estático		
	Equilibrado ( <i>balanced</i> )		
<b>Conteúdo</b>	Especializado	especializado (só comentários)	tipos específicos (em geral gêneros ou registros definidos)
	Regional ou dialetal		
	Multilíngüe		
<b>Autoria</b>	aprendiz	língua nativa	Os autores dos textos não são falantes nativos, não havendo identificação de falantes de outras línguas)
	língua nativa		

CRITÉRIO	TIPO	NETVIZZ/CORPUS ATUAL	EXPLICAÇÃO
Disposição interna	Paralelo	Não se aplica	
	Alinhado		
Finalidade	estudo	treinamento/teste	Foi construído para permitir o desenvolvimento de aplicações e ferramentas de análise.
	referência		
	treinamento ou teste		

**Quadro 2. Formação de *corpus* com Netvizz**  
 Fonte: o Autor (adaptado de Sardinha, 2000)

Como nosso objetivo, primeiro, aqui, é descrever o funcionamento das ferramentas, a representatividade do corpus é limitada: seu papel é servir de teste para mostrar como dispositivos computacionais analisam textos escritos em português brasileiro. Além disso, preferimos focar no gênero *comentários* porque, como dissemos na escolha do período, deve revelar um menor monitoramento linguístico por parte dos usuários, em grande parte, adeptos de polos opostos sobre a questão da prisão do ex-presidente. A seguir, antes da apresentação das demais ferramentas e da indicação da análise, exemplificamos o *corpus* com os 10 primeiros comentários dos 149 coletados e analisados aqui.

---

Vc está justificando violência ? No caso dos tiros está em investigação ,ja agressão de ontem é nítida de onde veio ... Mas vc está querendo justificar ... Ai é doença mesmo

---

As duas violências foram feitas por fanáticos covardes.

---

Os vídeos são incontestáveis. Não era um “opressor”. Foi um cidadão agredido de forma completamente desproporcional e violenta. Lamentável.

---

Quem mandou bater em todo mundo nos protestos de 2013?

---

Vitor Costa Isso justifica o quê? Guerra?

---

A esquerda está apanhando calada há muito tempo, está sendo vilipendiada em tudo. Todo mundo sabe que este indivíduo infelizmente foi lá provocar num momento muito sensível. Provocaram a esquerda demais, a esquerda vem sofrendo todo tipo de agressão e infelizmente vai piorar para o lado da esquerda mais uma vez.

---

A primeira coisa que eu pensei quando li este post ! Muito cinismo desta página!

---

Ai, lá vem o Cosimo Lowbike de novo, tá me perseguindo? Onde você leu eu defendendo violência? Me espantei com a chamada da reportagem falando de casos de violência dos últimos dias, como assim? E você Cosimo me viu falando contra a violência em outro post há uma hora atrás. Me poupe!

---



---

Injustificável! Mas como apareceu no JN.....soy contra. Fantoche é o c.!

---

Ricardo Oliveira..a esquerda é culpada por tudo isso ,aliás por onde a esquerda passa causa destruição etc .. Resultado ta ai ,vc mesmo justificando violência e se viabilizando etc ..

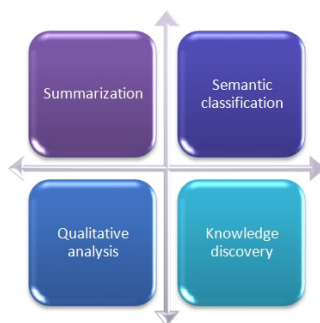
---

**Quadro 3. Apresentação de 10 dos 149 comentários analisados, entre os 1.150 coletados**  
**Fonte: Netvizz/Facebook.**

Com a coleta finalizada, o próximo passo é a análise dos dados. Para isso, vamos utilizar igualmente ferramentas computacionais que podem auxiliar o pesquisador, especialmente com *corpora* digitais.

### 3.2. Tropes: análise de dados

Criado nos anos 90, pela *Semantic-Knowledge* (Acetic), com sede em Paris, o *Tropes* é uma das ferramentas que a empresa desenvolveu para análise textual. Por meio da incorporação de conhecimentos advindos da área de Processamento de Linguagem Natural, o *software* foi criado servir áreas diversas, tais como os Sistemas de Informação, a Sociologia e a Linguística. O intuito da ferramenta focou-se em quatro aspectos na análise dos textos, apresentados na Figura 2, a seguir: resumo (*summarization*), classificação semântica (*semantic classification*), análise quantitativa (*quantitative analysis*) e descoberta de conhecimento (*knowledge discovery*).<sup>4</sup>



**Figura 2. Análises textuais possíveis com o Tropes**  
**Fonte: Tropes.**

---

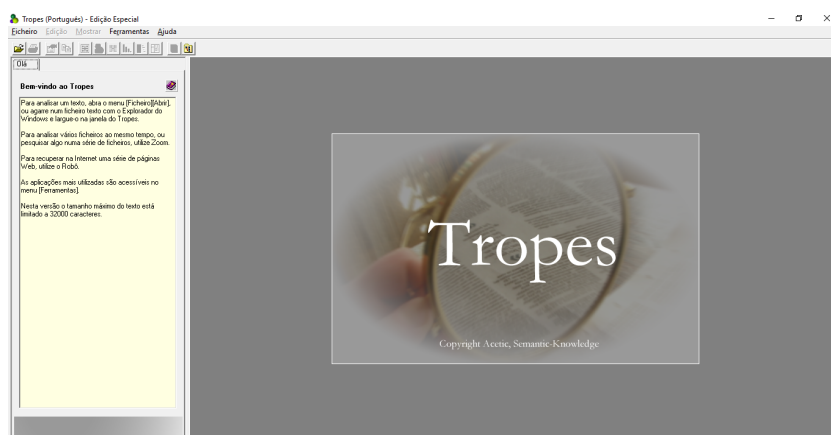
4 Mantivemos a figura em inglês pela fidelidade à fonte: embora o *software* tenha versão em português, a página está em inglês.

Tal como fizemos com o *Netvizz*, buscámos trabalhos no portal de periódicos da Capes e encontrámos apenas o trabalho de Araújo (2017) sobre o uso desse aplicativo em centros de pesquisa brasileiros. Ainda assim, o trabalho discutia a caracterização do gênero entrevista na língua espanhola. Em seu texto, Araújo (2017, p. 300) descreve que:

[este] *software* destaca-se pelo processamento semântico de textos em línguas naturais. Para descrever as características dos enunciados em análise, o Tropes 7.2.3 vale-se de critérios linguísticos pré-programados e os associa às estruturas linguísticas encontradas nos textos processados.

Por suas características de análise lexical, o programa faz um processamento refinado do qual decorrem informações tais como o “estilo textual”, o contexto básico relativo ao texto, denominado ali de “universo de referências” (sendo uma espécie de mapeamento do repertório mobilizado no texto); e dados quantitativos referentes a classes lexicais.

Para isso, é preciso que o pesquisador configure os documentos em formatação textual, preferencialmente em extensão de página web filtrada. Feito isso, ao acessar o aplicativo, o pesquisador poderá solicitar ao *software* o processamento dos textos salvos em uma pasta, individualmente ou de forma conjunta. A Figura 3 apresenta a tela de abertura da ferramenta.



**Figura 3. Tela inicial do Tropes**  
 Fonte: Tropes

Com a aplicação no *corpus* de treinamento aqui apresentado, a intenção era descrever suas características textuais a partir desse *software*. Assim,

elementos como “estilo textual”, “universo de referência” e algumas “categorias lexicais” foram selecionadas para apresentação neste trabalho. Embora a publicação contasse com 1.150 comentários, a restrição a 149 foi necessária devido ao número reduzido de caracteres para análise que possibilita o programa. O resultado é apresentado pela ferramenta em seções específicas. Começamos pelo estilo (Figura 4).



Figura 4. Estilo no Tropes  
Fonte: Tropes

Dos diferentes estilos categorizados no aplicativo (narrativo, descritivo, argumentativo e enunciativo), o *corpus* com 149 comentários foi analisado como predominantemente narrativo. Ainda que o esperado fosse um estilo argumentativo, percebe-se que o aplicativo justifica a escolha do estilo marcando no texto os elementos que o levaram a isso, entre eles, verbos de ação (em oposição a estativos), advérbios de tempo (“ontem” e “agora”, por exemplo), uso recorrente do conectivo “e”. Aqui, fica claro que essa predominância tem relação direta com o evento em si, denotando uma sequência de fatos que culminaram na prisão do ex-presidente.

Com relação ao “universo de referência”, o aplicativo apontou 28 agrupamentos de categorias de substantivos (referências), tais como “vida humana” (144 ocorrências), em itens como *violência* e *pessoas*; “conceitos gerais” (106 ocorrências), em casos como *esquerda* e *manifestações*; e “comunicação e mídia” (26 ocorrências), em exemplos como *vídeo* e *jornal*. Os casos sublinhados aqui apresentam uma noção do tema da publicação e dos eventos que estavam envolvidos no período da prisão, sobretudo aqueles envolvendo

manifestações contrárias e favoráveis a Lula, bem como episódios de violência que se registraram na ocasião. O resultado é apresentado na Figura 5.



Figura 5. Universo de referência no Tropes  
Fonte: Tropes

Em seguida, selecionamos as categorias lexicais com maior frequência: verbos factivos (61%); conectivo de adição (56,3%); modalização de negação (22,4%) e de tempo (21,8%); adjetivos subjetivos (56,1%); pronomes de primeira e segunda pessoa (15,1% e 24,5%, respectivamente). O pesquisador pode acessar todas as categorias analisadas pelo Tropes, assim como, ao clicar sobre cada uma, observar as expressões que a exemplificam no *corpus*. As Figuras 6 e 7, que se seguem, apresentam os dados completos.<sup>5</sup>

5 Embora as Figuras 6 e 7 apresentem muitos dados e pudessem ser apresentadas em tabelas, preferimos as figuras para que o leitor veja claramente como o aplicativo apresenta os textos analisados.

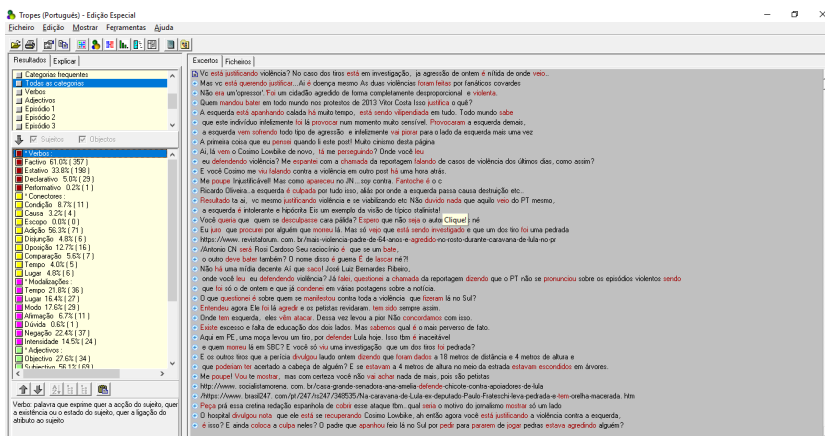


Figura 6. Categorias lexicais no Tropes – parte 1

Fonte: Tropes

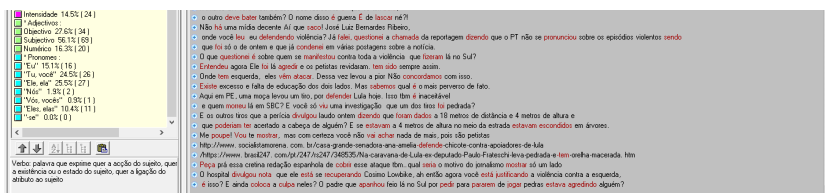


Figura 7. Categorias lexicais no Tropes – parte 2

Fonte: Tropes

Alguns destaques poderiam ser feitos nesse *corpus*, do ponto de vista linguístico: além do número produtivo de verbos factivos (de ação), o que representa bem o episódio da prisão, como dito antes, o baixo índice de pronomes representam a polaridade que tem dominado o cenário de discussões políticas no País. Por outro lado, a vasta ocorrência de itens de conexão e modalização fornecem dados importantes quanto às estratégias textuais de coesão e posicionamento nos comentários *on-line*, que poderiam ser comparados com *corpora* distintos. Tudo isso justifica o trabalho de coleta e análise por meio de ferramentas computacionais.

A seguir, debruçamo-nos sobre o *Linguakit* e mostramos de que forma este portal pode contribuir para a análise linguística.

### 3.3. Linguakit

Desenvolvido pelo *Cilenis Language Technology*, da Universidade de Santiago de Compostela, o *Linguakit* é um *site* multilíngue com diversas ferramentas de uso linguístico, baseadas em Processamento de Linguagem Natural, tais como resumidor, analisador de sentimentos ou de frequência de palavras entre muitas outras (Figura 8). A maior parte dessas ferramentas é de uso gratuito e apresenta resultados interessantes no que diz respeito à análise de dados.

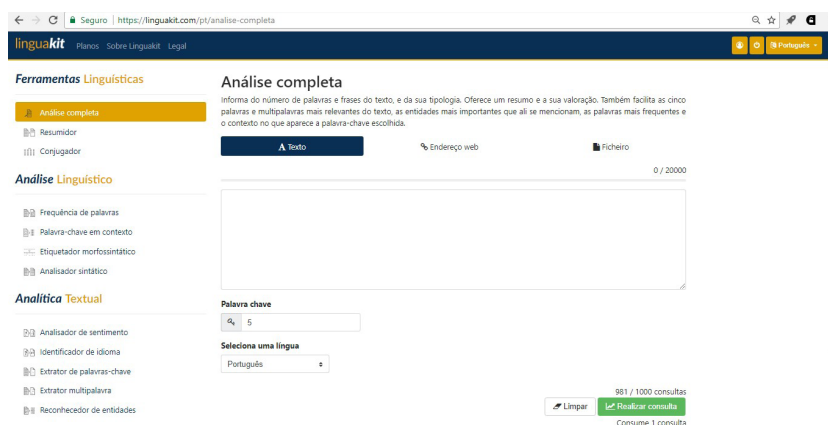


Figura 8. Página inicial do Linguakit  
Fonte: Linguakit.

Para apresentá-la, neste trabalho, tomamos 149 textos da postagem mais comentada no *El País Brasil*, entre 7 e 8 de abril de 2018, conforme descrevemos na subseção anterior. Embora a publicação tivesse 1.150 comentários, foi feita uma seleção de 149 por restrição do programa (conferir a primeira nota de rodapé). Começamos a descrição da análise do *Linguakit* pelo extrator de palavras-chave (Figura 9).



Figura 9. Nuvem de palavras do Linguakit

Fonte: O autor.

Nesse caso, o resultado nos permite observar os elementos mais proeminentes nos 149 comentários: percebemos que “Lula”, “petistas” e “selvageria” foram os elementos mais recorrentes nesse *corpus*. Assim, o extrator aponta para o contexto sobre o qual os comentários versavam.

Cabe observar que itens como “q” ou “vc” foram igualmente bem citados nos textos; nesse sentido, vale destacar que o aplicativo pode contribuir para pesquisas sobre elementos próprios da linguagem *on-line*, já que mostra sua frequência. Como defendem alguns autores (Recuero, 2012; Coulmas, 2014), a Comunicação Mediada por Computador (CMC) é uma nova e importante área de pesquisa sobre as modificações linguísticas que ali podem ocorrer (tanto na fala como na escrita): não só uma oralização é comum (uso de emoticons ou repetição de letras na tentativa de indicar a entonação ou ação, por exemplo), como a presença de abreviações e inovações podem ser descritas. A tabela a seguir mostra a frequência de alguns desses itens.

**Tabela 1. Frequência de itens CMC do Linguakit**

<b>frequência</b>	<b>item</b>
13	vc
7	q
4	tbm
4	vcs
4	post
3	pq
2	tb
1	uhhhhh
1	uééééé
1	heim
1	kkkkkkk

**Fonte: O autor.**

O resultado acima nos leva a algumas conclusões: primeiro, que a abreviação parece ser o tipo de item da CMC mais recorrente; depois, que entre elas, “vc” e “q” estão praticamente estabelecidas. Naturalmente, isso deve ser confrontado com outros dados, especialmente levando em conta os tipos de páginas e postagens realizadas na rede social. De qualquer modo, conseguimos mostrar que a ferramenta aponta para elementos importantes no que diz respeito a essa estratégia de escrita em ambiente digital. Sem dúvida, tanto essas questões como aquelas apontadas pelo Tropes podem servir para descrição do gênero *comentário on-line*, sendo igualmente possível a busca por elementos de variação, especialmente na escrita, tal como nos apontam Paredes e Silva (2012) ser um ponto fundamental nas pesquisas da área.

O último item que desejamos apresentar no *Linguakit* é sua ferramenta de análise de sentimentos. Segundo Benevenuto *et al.* (2015, p. 2),

o principal objetivo da análise de sentimentos é definir técnicas automáticas capazes de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão.

Para fazer a classificação, o software é programado para medir a polaridade da frase, sendo ela classificada, no *Linguakit*, de forma ternária, a



saber: positiva, negativa ou neutra. Novamente, Benevenuto *et al.* (2015, p. 3) nos explicam a diferença:

(...) a frase “Como você está bonita hoje” é *positiva* e a frase “Hoje é um péssimo dia” é *negativa*, já a frase “Hoje é 21 de Outubro” não possui polaridade e normalmente é classificada como *neutra*. (grifos dos autores)

Nesse sentido, a polaridade tem relação direta com os itens que, na sentença, podem ser mais associados com questões subjetivas, apontadas, por exemplo, por adjetivos.

Considerando o tema dos comentários analisados no presente trabalho, bem como os resultados prévios a respeito das palavras mais frequentes no contexto, poderíamos levantar a hipótese de que os comentários tenderão a se encontrar no polo negativo, já que palavras como “selvageria”, “violência” e “bandido” apareceram com frequência na nuvem de palavras. A Figura 10, a seguir, apresenta os resultados do *Linguakit*.

### Sentimento do texto

#### Estatística

Frases negativas	Frases neutras	Frases positivas
83	43	23

#### Sentimento por frase

Vc está justificando violência ? No caso dos tiros está em investigação ,ja agressão de ontem é nítida de onde veio .. Mas vc está querendo justificar ... Ai é doença mesmo	100.00%	Positivo
As duas violências foram feitas por fanáticos covardes.	-100.00%	Negativo
Os vídeos são incontestáveis. Não era um "opressor". Foi um cidadão agredido de forma completamente desproporcional e violenta. Lamentável.	-100.00%	Negativo
Quem mandou bater em todo mundo nos protestos de 2013?	-93.85%	Negativo
Vitor Costa Isso justifica o quê? Guerra?	-94.71%	Negativo

Figura 10. Análise de sentimento do *Linguakit*

Fonte: O autor

De fato, como se esperava, houve uma predominância de frases analisadas como negativas no contexto dos 149 comentários selecionados no *El País Brasil*. Embora essa área seja menos explorada por linguistas, não deixa de ser importante chamar a atenção para essa funcionalidade do *Linguakit*.

Nesta seção, apresentamos algumas das principais funcionalidades do *Tropes* e do *Linguakit*, ferramentas de análise textual digitais que podem contribuir para o trabalho dos linguistas.

#### 4. Considerações finais

Neste trabalho, descrevemos tomamos um corpus de rede social (*Facebook*), por meio do aplicativo *Netvizz* nela integrado e o submetemos a uma análise automática de textos, por meio dos aplicativos *Tropes* e *Linguakit*. O primeiro software indicou questões textuais importantes, como o estilo geral do corpus, o contexto de referência e as categorias frequentes. Com tais dados, sugerimos que pesquisas sobre tais tópicos poderiam ser facilitadas com a ferramenta, que faz todo o trabalho de mapear as ocorrências no texto. A segunda ferramenta apontou para análises importantes, como o índice de palavras-chave, que indicam a frequência mais alta de termos no corpus, como também a ocorrência de termos próprios da CMC (abreviações, inovações e repetições de letras). Finalmente, mostramos a funcionalidade “análise de sentimentos” e, a partir das questões lexicais apontadas pelas ferramentas, ligadas a violência, sugerimos que os comentários seriam analisados, em sua maioria, como negativos, o que de fato ocorreu: o aplicativo apontou um total de 83 frases negativas, 43 neutras e apenas 23 positivas.

Em suma, as análises apresentadas têm como intenção chamar a atenção de pesquisadores para novas estratégias de coleta e análise de dados, especialmente em redes sociais. Sua importância vem sendo discutida por diferentes autores. Sendo um espaço de engajamento, é ali também que ocorrem manifestações linguísticas relevantes. Nesse sentido, Benevenuto *et al.* (2015, p. 2) argumentam que

as redes sociais são a criação de uma revolução digital, permitindo a expressão e difusão das emoções e opiniões através da rede. De fato, redes sociais são locais onde as pessoas discutem sobre tudo expressando opiniões políticas, religiosas ou mesmo sobre marcas, produtos e serviços.

Para os linguistas, dedicados ao estudo da linguagem em suas diferentes manifestações, parece fundamental esse olhar voltado a um ambiente tão importante como o digital. Esperamos que este trabalho tenha contribuído para isso.

## Referências

- Araújo, J. Leffa, V. (Ed.) (2016). *Redes sociais e ensino de línguas: o que temos de aprender?*. (1ª ed.) São Paulo: Parábola Editorial.
- Araújo, L. S. de. (2017). O gênero entrevista radiofônica em comunidades hispânicas: um aporte da Análise Textual Automática. *Domínios de Linguagem*, 11(2), 289–312. Disponível em: DOI: 10.14393/DL29-v11n2a2017-2. Consultado em: 23. abr. 2018.
- Auroux, S. (2014). *A revolução tecnológica da gramatização*. (3ª ed.) Campinas, Brasil: Editora da UNICAMP.
- Barton, D. & Lee, C. (2015). *Linguagem online: Textos e práticas digitais*. M. C. Mota (Trad.). (1ª ed.) São Paulo: Parábola Editorial.
- Benevenuto, F., Ribeiro, F. & Araújo, M. (2015). *Métodos para Análise de Sentimentos em Mídias Sociais*. Short course in the Brazilian Symposium on Multimedia and the Web (Webmedia). Manaus. Disponível em: <<http://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>>. Consultado em: 26 abr. 2018.
- Bertucci, R. A. & Nunes, P. A. (2017). Interação em rede social: Das reações às características do gênero comentário. *Domínios de Linguagem*, 11(2), 313–338. DOI: 10.14393/DL29-v11n2a2017-3.
- Cassirer, E. (1979). *Antropologia filosófica: Ensaio sobre o homem*. São Paulo: Mestre Jou.
- Coscarelli, C. V. (2016). Navegar e ler na rota do aprender. In C. V. Coscarelli (Ed.), *Tecnologias para aprender*. São Paulo: Parábola.
- Coulmas, F. *Escrita e Sociedade*. (2014). São Paulo: Parábola Editorial.
- Cupani, A. (2016). *Filosofia da tecnologia: um convite*. Florianópolis: Editora da UFSC.
- Finatto, M. J. B. (2017). Descrição de gêneros textuais/discursivos com apoio computacional. *Domínios de Linguagem*, 11(2), 282–288. doi: 10.14393/DL29-v11n2a2017-1. Linguakit. <http://linguakit.com/pt/>. Consultado em: 26 abr. 2018
- Manifesto das Humanidades Digitais. THATCamp Paris. 2012. Disponível em: <<https://humanidadesdigitais.org/manifesto-das-humanidades-digitais>> Consultado em: 26 abr. 2018.
- Netvizz. <https://apps.facebook.com/netvizz/>. Consultado em: 26 abr. 2018
- Ortega y Gasset, J. (1965). *Meditación de la técnica*. Madrid: Espasa-Calpe, (orig. 1939).
- Paiva, C. & Paredes-Silva, V. L. (2012). Cumprindo uma pauta de trabalho: Contribuições recentes do PEUL. *Alfa*, São Paulo, 56(3), 739–770.
- Portal de Periódicos da Capes. [www.periodicos.capes.gov.br](http://www.periodicos.capes.gov.br). Consultado em: 26 abr. 2018.
- Raso, T. & Mello, H. (Eds.) (2012). *C-ORAL-BRASIL I: corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG.
- Recuero, R. (2012). *A conversação em rede: comunicação mediada pelo computador e redes sociais na Internet*. Porto Alegre: Sulina.

- Rieder, B. (2013). Studying Facebook via data extraction: The Netvizz application. *Proceedings of the 5th Annual ACM Web Science Conference*, 346–355.
- Sardinha, T. B. (2000). Corpus Linguistics: History and problematization. *DELTA*, 16(2), 323–367. Disponível em: <<http://dx.doi.org/10.1590/S0102-44502000000200005>> Consultado em: 26 abr. 2018.
- Sardinha, T. B. (2005). Trazendo a língua portuguesa para o computador. In T. B. Sardinha (Ed.), *A Língua Portuguesa no computador* (pp. 269–295). Campinas: Mercado de Letras & Fapesp.
- Tropes. <http://www.semantic-knowledge.com/tropes.htm>. Consultado em: 26 abr. 2018.
- Vieira Pinto, Á. (2005). *O conceito de tecnologia*. Rio de Janeiro: Contraponto.

[recebido em 26 de abril de 2018 e aceite para publicação em 30 de março de 2019]



# ANÁLISE DIACRÓNICA DOS TEMPOS COMPOSTOS *TINHA FEITO, TEREI FEITO E TERIA FEITO* NA LÍNGUA PORTUGUESA

DIACHRONIC ANALYSIS OF COMPOUND TIMES IN  
PORTUGUESE: *TINHA FEITO, TEREI FEITO AND TERIA FEITO*

Jan Hricsina\*

jan.hricsina@ff.cuni.cz

Este artigo foca a análise diacrónica dos tempos compostos *tinha feito, terei feito e teria feito* na língua portuguesa. O seu objetivo principal é analisar as funções modo-temporais dos tempos compostos em questão no Português Antigo e comparar a sua frequência e o seu uso na evolução da língua portuguesa. A pesquisa tem por base o corpus linguístico [www.corpusdoportugues.org](http://www.corpusdoportugues.org).

**Palavras-chave:** Língua portuguesa. Linguística diacrónica. Linguística de *corpus*. Linguística funcional. Tempos compostos *tinha feito, terei feito e teria feito*.

The paper focuses on the diachronic analysis of the compound tenses *tinha feito, terei feito* and *teria feito* in the Portuguese language. The principal objective of this study is to analyze the modo-temporal functions of these compound tenses in Old Portuguese, to compare its frequency and its use in the evolution of the Portuguese language. The research is based on the linguistic *corpus* [www.corpusdoportugues.org](http://www.corpusdoportugues.org).

**Keywords:** Portuguese language. Diachronic linguistics. *Corpus* linguistics. Functional linguistics. Compound tenses *tinha feito, terei feito* and *teria feito*.



## 1. Introdução

Do ponto de vista da tipologia morfológica, podemos constatar que um dos fenómenos panromânicos, ou seja, comuns à evolução de todas as línguas românicas, é representado pela tendência analítica que se manifesta

---

\* Universidade Carolina – Univerzita Karlova, República Checa.

em vários subsistemas desta natureza. Entre as mudanças mais notórias, podemos mencionar a formação do artigo, o uso frequente de preposições ou a gradação de adjetivos e advérbios por meio de palavras auxiliares (*mais* e *menos*). No subsistema verbal, a tendência analítica manifesta-se na formação de perífrases verbais que podem tornar-se tempos compostos. Estas estruturas analíticas completam ou enriquecem o subsistema verbal. Formam, assim, uma perspetiva temporal secundária.<sup>1</sup>

Distinguem-se geralmente dois tipos de formação de perífrases verbais. O primeiro tem uma estrutura formada pelo verbo auxiliar *habere* ou *tenere* e participio passado. Estes paradigmas verbais<sup>2</sup> denotam tradicionalmente acontecimentos ou eventos ocorridos num momento anterior relativamente ao ponto de referência<sup>3</sup> delimitado pelo tempo do verbo auxiliar. Por exemplo, o tempo composto *tinha feito* expressa situações anteriores ao ponto de referência representado pelo tempo do auxiliar *tinha*, ou seja, que se encontra no passado; *P-terei feito* pode denotar situações anteriores a outros acontecimentos futuros que representam o ponto de referência deste tempo e assim por diante.<sup>4</sup> O segundo tipo é representado pelas perífrases formadas pelo verbo semi-auxiliar *ir* e infinitivo.<sup>5</sup> Estes paradigmas exprimem acontecimentos que são posteriores ao ponto de referência representado pelo tempo do verbo semi-auxiliar.<sup>6</sup> Consideremos o exemplo de P-vou fazer que denota

- 
- 1 O eixo desta perspetiva temporal é formado por duas relações temporais: 1. posterioridade relativa; 2. anterioridade relativa. O futuro composto (*terei feito*) denota processos anteriores relativamente a outros acontecimentos futuros. Este tempo pertence, assim, à perspetiva temporal secundária (Zavadil; Čermák 2010, p. 274 *apud* E. Coseriu, *Das romanische Verbalsystem* 1976).
  - 2 Por termo paradigma verbal entendemos o conjunto de todas as formas gramaticais de um verbo que têm o mesmo valor modo-temporal. Neste artigo, para este termo vamos usar a abreviatura P.
  - 3 “O ponto de referência serve como ponto intermédio a partir do qual se pode situar o evento (ou estado) descrito.” (Oliveira 2004, p. 131).
  - 4 Nenhum dos tempos compostos deste tipo exprime o valor temporal de anterioridade. O pretérito perfeito composto português (*tenho feito*) exprime uma ação durativa ou reiterativa que começou num momento do passado (geralmente, não sabemos exatamente quando) e que dura até ao momento presente e há possibilidade de esta situação ou estado se prolongar até ao futuro. Também o pretérito perfeito composto espanhol nem sempre denota uma ação passada simples (Zavadil & Čermák 2010, pp. 275–277). Por outro lado, o valor temporal de anterioridade não é a única função que podem adquirir os tempos compostos deste tipo. Em Português, *P-terei feito* e *P-tinha feito* podem ter também leituras modais (ver mais adiante).
  - 5 Em Espanhol, a perífrase é formada pelo verbo semi-auxiliar *ir* + preposição *a* + infinitivo.
  - 6 Do ponto de vista semântico, o verbo *ir* deve ser considerado semi-auxiliar devido ao facto de poder conservar o seu próprio significado lexical (deslocar-se de um lugar a outro) em certos contextos (*Vou ver um amigo*.) (Tláškal 1978, p. 205). Do ponto de vista sintático, existem vários testes sintáticos que servem a classificar um verbo seja como auxiliar seja como semi-auxiliar (Paiva Raposo 2013, pp. 1238–1256).

situações posteriores ao ponto de referência (delimitado pelo tempo do verbo semi-auxiliar) que é simultâneo ao momento da fala.<sup>7</sup> Outros tipos de estruturas analíticas verbais formadas por vários verbos semi-auxiliares (*ir*, *vir*, *começar*, *acabar*) e gerúndio ou por estes verbos, preposição e infinitivo veiculam valores aspetuais. Neste trabalho, não nos ocuparemos delas.

No presente artigo vamos analisar o primeiro tipo de estruturas analíticas, ou seja, os assim chamados tempos compostos, mais precisamente: *tinha feito*, *terei feito* e *teria feito*.<sup>8</sup> Vamos observar o emprego e a frequência destes paradigmas na evolução da língua portuguesa, centrando-nos nos seus valores modo-temporais. Para tal, vamos aproveitar o *corpus* linguístico que permite fazer pesquisas diacrónicas [www.corpusdoportugues.org](http://www.corpusdoportugues.org).<sup>9</sup>

## 2. Os tempos compostos no Português contemporâneo

Nesta parte do trabalho, analisaremos o emprego dos tempos compostos referidos, no Português contemporâneo.

### 2.1. P-tinha feito

A função principal deste paradigma é exprimir processos que ocorreram anteriormente ao ponto de referência que se encontra no passado. O ponto de referência pode ser expresso por outro tempo (*O Paulo disse que a Maria tinha apanhado muito trânsito a caminho do trabalho*.<sup>10</sup>) ou por outra frase (*A Maria chegou atrasada ao trabalho. Tinha apanhado muito trânsito*.) (Oliveira 2013, pp. 530–531; cf. também Svobodová 2014, pp. 83–84). Pode também acontecer que o ponto de referência não figure no texto ou na comunicação, sendo expresso implicitamente (deduzível do contexto). Veja-se a frase seguinte: *O Braga tinha ganho muitos jogos*. Esta frase sugere a ideia de que o Braga tinha ganho muitos jogos dantes e ganhou

7 O momento da fala representa o segmento temporal em que uma frase é proferida (cf. Oliveira 2013, p. 510–511).

8 Neste estudo, não incluímos P-tenho feito visto que já o analisámos noutro lugar (Hricsina 2017).

9 O corpus elaborado por Mark Davies (BYU) e Michael J. Ferreira (Georgetown University) contém mais de 45 milhões de palavras nos textos provenientes dos séculos XIII–XX escritos em ambas as variantes principais do Português, respetivamente no Português Europeu e no do Brasil.

10 Caso não seja indicada a fonte, os exemplos são do autor.



novamente, representando esta última vitória o ponto de referência. Cunha e Cintra referem que este paradigma pode ser usado também para atenuar uma afirmação ou pedido, ou seja, para exprimir um facto passado em relação ao momento presente – *Eu tinha vindo para convencê-lo de que Pedro é seu amigo e pedir-lhe que apoiasse Hermeto* (Cunha & Cintra 1999, p. 455).

Atualmente este paradigma pode figurar também em orações condicionais contrafactuais, substituindo *P-teria feito*. Nestas construções pode ter um valor modo-temporal, sendo uma variante preferida pelos falantes (Svobodová 2014, p. 100).

## 2.2. P-terei feito

Este paradigma expressa um acontecimento posterior ao tempo da fala e anterior ao ponto de referência que se encontra no futuro. O ponto de referência pode ser delimitado quer por um sintagma adverbial, quer por outra frase (*Daqui a uma semana, o trabalho terá sido feito. – Quando chegarmos a casa, o Pedro terá preparado o jantar.*). Neste emprego, *P-terei feito* encontra-se muitas vezes substituído por *P-fiz*. Muitos falantes preferem usar este paradigma visto que diferentemente de *P-terei feito*, este carece de informação modal (ver mais adiante) (Oliveira 2013, pp. 531–532).

Quando *P-terei feito* indica um processo ocorrido anteriormente ao tempo da fala, tem uma leitura modal, exprimindo uma incerteza, probabilidade ou não comprometimento do falante com a situação representada<sup>11</sup> (*O acidente terá acontecido mesmo assim?*) (Oliveira 2013, p. 532; Cunha & Cintra 1999, p. 460). Svobodová acrescenta que este paradigma adquire o valor modal apenas quando denota ações concluídas ou estados (Svobodová 2014, p. 93).

Celso Cunha e Lindley Cintra (1999, p. 460) afirmam que *P-terei feito* serve também para exprimir uma ação futura certa. Vejamos o exemplo referido na gramática destes autores: *Só o Direito perdurará e não terá sido vão o esforço da minha vida inteira.*

Duarte (2009) afirma que, no Português contemporâneo, o futuro composto se usa sobretudo no discurso jornalístico e o seu emprego é típico das frases simples e com valor modal. Acrescenta que se trata de um tempo verbal de relato, ou seja, alguém relata um facto que soube de outra pessoa (o mediativo), não se responsabilizando pela validação do conteúdo da informação.

---

11 Quando o ponto de referência coincide com o momento da fala, o paradigma denota um acontecimento provável ou incerto e temporalmente ligeiramente anterior ao momento da fala (*Neste momento, o avião já terá aterrado no Porto.*) (Oliveira 2013, p. 532).

### 2.3. P-teria feito

Este paradigma denota situações anteriores ao ponto de referência que se encontra no passado e é caracterizado por uma forte modalização. Estes processos são considerados como incertos ou prováveis (Oliveira 2013, p. 532) (*Ontem não encontrei o professor Rodrigues na faculdade; teria ido a uma conferência.*)

P-teria feito pode adquirir também uma leitura contrafactual, ou seja, pode denotar uma situação (anterior ao ponto de referência passado) dependente de uma condição. A possibilidade de se concretizar esta situação não se realizou. A condição pode ser expressa seja, por exemplo, por um sintagma preposicional seja por uma frase condicional (Oliveira 2013, p. 533; Cunha & Cintra 1999, p. 463). (*Com a Maria, o Pedro teria tido uma vida mais feliz. / Se tivesse entregado o TPC a tempo, não teria tido problemas com o exame.*)

No Português contemporâneo, este paradigma passou a ser uma forma pouco usada na língua falada. Tende, cada vez mais, a ser substituído e com uma maior frequência por P-faria ou P-tinha feito (em orações condicionais) (Svobodová 2014, pp. 99-100) (*Se o Paulo tivesse concordado, tínhamos vendido a casa.*) Porém, caso P-tinha feito não figure em uma oração condicional, perde o seu valor de condicional (*Ontem não encontrei a nossa vizinha. Teria/tinha partido para Évora.*).

## 3. Os tempos compostos na história da língua portuguesa

Os especialistas na evolução do Português confirmam a existência de tempos compostos já na fase inicial do Português escrito, ou seja, no século XIII (Mattos & Silva 2008, p. 441). Neste período, a frequência de tempos compostos é, porém, muito baixa (Huber 2006, p. 252-254).<sup>12</sup> Acrescente-se que nem todas as estruturas constituídas pelos verbos *ter*, *haver* e *ser*<sup>13</sup> representavam

12 A frequência de P-tinha feito começou a crescer consideravelmente desde o século XV. Aproximadamente neste período ocorreu a convergência das vogais ou ditongos nasais finais no ditongo universal /ẽw̃/. Este fenómeno afetou provavelmente o funcionamento do mais-que-perfeito simples visto que a terceira pessoa do plural deste tempo passou a ser idêntica à mesma forma de P-fiz (*fizeram*). Foi muito provavelmente este motivo que levou à propagação da forma composta do mais-que-perfeito (Brocardo 2014, pp. 154-156).

13 Durante a evolução da língua portuguesa podemos considerar três verbos auxiliares: *haver*, *ser* e *ter*. O verbo *haver* foi usado até ao século XV e desde aí foi sendo substituído pelo verbo *ter*. O auxiliar *ser* nunca foi tão frequente, sendo o seu emprego limitado aos verbos de movimento (*vir*, *partir*, *chegar*, *ir*) ou intransitivos (*falecer*). Deixou de ser empregue como auxiliar só no século XIX (Hricsina 2017, p. 182).

tempos compostos. Além de tempos compostos, os sintagmas formados por um dos verbos mencionados e um particípio podem ter representado também estruturas transitivas predicativas (sobretudo com concordância do particípio) (Brocardo 2014, p. 152). Naquela altura, este tipo de estruturas ainda não foi gramaticalizado como tempo composto. A sua gramaticalização terá ocorrido provavelmente da seguinte maneira: nos séculos XIV e XV, a estrutura formada pelo verbo *haver* e particípio representava um tempo composto, enquanto a estrutura com o auxiliar *ter* configurava uma construção resultativa. Em ambas as estruturas encontramos frequentemente a concordância do particípio com o objeto direto.<sup>14</sup> No século XVI, a frequência do verbo auxiliar *haver* começa a diminuir, sendo substituído pelo auxiliar *ter* em ambas as construções (tempo composto e construção resultativa). Enquanto que no tempo composto, a concordância participial se torna cada vez mais esporádica, na construção resultativa a concordância passa a ser obrigatória (Hricsina 2017, p. 182-183).<sup>15</sup> É aproximadamente neste período que o tempo composto se gramaticaliza.<sup>16</sup>

#### 4. Métodos de análise

Para analisar a frequência e o emprego dos tempos compostos (P-*tinha feito*, terei feito e teria feito) na evolução da língua portuguesa, servimo-nos do *corpus* linguístico disponível em [www.corpusdoportugues.org](http://www.corpusdoportugues.org), que permite fazer pesquisas diacrónicas. Procurámos todas as ocorrências dos tempos compostos em questão (formados por vários verbos auxiliares) (\_vi\* \_vk\*; \_vf\* \_vk\*; \_vc\* \_vk\*) entre os séculos XIII e XX.<sup>17</sup> De todas as ocorrências encontradas no *corpus* analisámos 100 casos de cada tipo de tempo composto para cada século, sempre que tal número de exemplos estava disponível. A seleção dos casos analisados foi aleatória.

14 Considerem-se estes exemplos:

1) *E dhy partyo logo pera Bizcaya, que tiinha prometida ao iffante dō Johā, seu primo.* (Crónica Geral de Espanha de 1344)

2) *Em quanto el rei tiinha cercada esta cidade, acaeceu que hũu bispo de Grecia leixara o bispado por mais livremente servyr a Deus e veuo ã romaria a Santiago.* (Crónica Geral de Espanha de 1344).

No primeiro exemplo, a construção *tiinha prometida* é um tempo composto e no segundo, trata-se de uma construção resultativa (Hricsina 2017, p. 172).

15 Segundo Tibor Berta, este facto é muito importante para os locutores poderem distinguir entre tempo composto (não-concordância) e construção resultativa (concordância) (Berta 2016).

16 A gramaticalização exhibe várias fases: 1. verbo pleno; 2. construção predicativa; 3. forma perifrástica; 4. aglutinação (Ribeiro 1996, p. 346). Podemos, assim, constatar que, no século XVI, os tempos compostos portugueses entraram na terceira fase da gramaticalização.

17 A análise *in corpora* foi feita em janeiro e fevereiro de 2018.

## 5. Análise em *corpora*

### 5.1. Século XIII

No que diz respeito a *P-tinha feito*, encontrámos, na totalidade, 41 ocorrências. Em todos os casos, este paradigma denota situações anteriores ao ponto de referência, que se encontra no passado.

- 1) *Esta é como Santa Maria levou o boi do aldeão de Segovia que ll' avia prometudo e non llo queria dar.* (Cantigas de Santa Maria 1)
- 2) *E com dereito seer enforcado deve Dom Pedro, porque foi filhar V15 a Cotom, poi'lo houve soterrado, seus cantares, e nom quis en de dar um soldo pera sa alma quitar sequer do que lhi havia emprestado.* (Cantigas de Escárnio e Maldizer)

Quanto a *P-terei feito*, no *subcorpus* do século XIII, encontrámos apenas uma ocorrência. Neste caso, este paradigma expressa uma ação posterior ao tempo da fala com um valor modal muito forte visto que a frase começa pelo verbo *creer* (não é detetável nenhum ponto de referência que se encontre no futuro) (ex. 3).

- 3) *Como Santa Maria fez soltar o ome que andara gran tempo escomungado. A creer devemos que\* todo pecado Deus pola sa Madr' averá perdôado.* (Cantigas de Santa Maria 1)

*P-teria feito* não foi encontrado no *subcorpus* do século XIII.

### 5.2. Século XIV

No *subcorpus* do século XIV, *P-tinha feito* está representado por 263 ocorrências, na totalidade. Nos 100 casos analisados, este paradigma denota sempre uma situação anterior ao ponto de referência, que se situa no passado (ex. 4 e 5).

- 4) *El rey dō Fernando era homē de b õõ talante e pesoulhe muyto do mal que avya recebido de seu irmão; e com piedade e mesura nõ quis a ello tornar.* (Crónica Geral de Espanha de 1344)
- 5) *E esto por que os homēes sabyam certamente que el rey avya jurado que nu n ca se levantarya de sobr'ella ataa que a tomasse.* (Crónica Geral de Espanha de 1344)

Quanto a *P-terei feito*, encontrámos apenas uma ocorrência, em que o paradigma se encontra entre o tempo composto e uma construção resultativa, exprimindo um processo culminado ou estado posterior ao tempo da fala (ex. 6).

- 6) *Ssenhor ante que al ffaçamos trabalhemonos pera prender aquel maaõ homẽ barllaao E sseo podermos tomar **aueremos acabado** todo nosso ffeyto. ca lhe ffaremos pora ffaagos ou portormentos que el coffesse que todo aquello que elle essynou aoteu ffilho eram cousas ffallssas e de grande erro.* (Barlaam e Josephat – 1967)

No *subcorpus* do século XIV, não foi encontrada nenhuma ocorrência de *P-teria feito*.

### 5.3. Século XV

No *subcorpus* do século XV, *P-tinha feito* está representado por 205 ocorrências. Nos 100 casos analisados, este paradigma exprime processos ocorridos anteriormente ao ponto de referência, que se encontra no passado (ex. 7 e 8).

- 7) *E sentindo-se assi çarrado tiinha o seu coração bem temeroso e cheo de maa vootade, nembrando-se em como Potem e Achilas matarom Pompeeo que lhes **avia feito** tanto bem, e temia-se de lhe fazerem outro tal, e peor, se caisse em seu poder, por que bem sabia que moor odio lhe aviam que a esse Pompeeo, e ele nom podia seer acorrido de sua gente.* (Vida e feitos de Júlio César)
- 8) *Tanto que ao condestabre, a Castelo Branco, honde estava, veeo recado que o iffante dom Diinis **era tornado** pera Castella, hordenou pera se hiiir a ehrey a Tuuy, como avia seu mandado.* (Estoria de Dom Nuno Alvares Pereyra)

Quanto a *P-terei feito*, tal como no século anterior, a sua frequência é muito baixa. No *subcorpus* respetivo, encontrámos apenas dois casos. Ao analisar estas duas ocorrências que surgem em uma só frase, pudemos constatar que o emprego deste paradigma se encontrava longe de estar estabilizado. No primeiro caso, o paradigma em questão expressa uma ação ocorrida após o ponto de referência no futuro e, no outro, denota um processo culminado (ou construção resultativa) anteriormente ao mesmo ponto de referência (ex. 9).

- 9) *E nestas tigeladas d'aRoz quẽ quer lhe dejta por cima ggemas d'ovos ẽtejras / beilhos d'aRoz Receita. depois que o aRoz estiuer cozido com leyte & temperado como ha d'estar & fryo / **tereis batidos** dous ovos cõ huã colher de farynha / & tomareis do aRoz que ja estara Frjo & deitaloeis nestes ovos que ja **tereis batidos**.* (Tratado de cozinha portuguesa)

No que diz respeito a *P-teria feito*, encontrámos 4 ocorrências (um paradigma formado pelo auxiliar *ter* e três com o *haver*). Em todos os casos, o paradigma em questão exprime situações incertas ou eventuais ocorridas anteriormente ao ponto de referência, que se encontra no passado (ex. 10).

- 10) *O comde mamdou que ho aguardassẽ em çima do porto, em hũa sellada que se ally faz, & que mãdassẽ estar allem de sy os allmogavares. & des que emtemdeo que **teriam passado** o mao caminho, de guisa que a mestura dos cavallos nõ podessẽ empeçer aos de pee, partio da çidade.* (Gomes Eanes de Zurara, Crónica do Conde D. Pedro de Meneses)

#### 5.4. Século XVI

No *subcorpus* do século XVI, encontrámos 1 870 ocorrências de *P-tinha feito* e, tal como nos séculos anteriores, este paradigma denota exclusivamente processos ocorridos anteriormente ao ponto de referência, que se encontra no passado (ex. 11).

- 11) *Era este Malec Caez vassalo do Rei da Pérsia, e tinha-lhe mandado pedir socorro contra o inimigo; e quando lhe chegou, já **tinha perdido** o Estado.* (João de Barros, Quinta década (livros 8–10), vol. 1)

Quanto a *P-terei feito*, no *corpus* respetivo, encontrámos apenas 2 ocorrências. Em ambos os casos, o paradigma exprime uma ação futura (com um valor modal forte), numa subordinada completiva cujo verbo da oração principal é *crer* (ex. 12 e 13).

- 12) *Eu, ellRey, vos emvio muito saudar. Bem creo que **teereis sabido** da vinda de Pero Lopez de Souza, que veyo do Brasill; o quall, antre outras boas novas que trouxe, foy que, vymdo elle do Rio da Prata, correndo a costa do Brasill, veyo teer a Pernambuco, õde achou os Franceses, que tinham feyto fortalleza;* (D. João III., *Letters of John III – King of Portugal 1521-1557*)

- 13) *Eu, elRey, vos envio muito saudar. Bem creio que **teereis sabido** como Amtonio de Brito se foy, sem se espedir de mim mî. (D. João III., Letters of John III – King of Portugal 1521-1557)*

No subcorpus do século XVI, P-teria feito encontra-se representado por 21 ocorrências. Em todos os casos, exprime situações incertas ou eventuais, e ocorridas anteriormente ao ponto de referência passado (ex. 14 e 15).

- 14) *E, parecendo ao Infante que já **teria sabido** muitas, porque o espírito o não deixava assossegat nestas que desejava saber daquelas partes, tornou a mandar o mesmo Antão Gonçalves em busca dele, e em sua companhia foram Garcia Mendes e Diogo Afonso, cada um em sua caravela. (João de Barros, Décadas da Asia (Década Primeira, Livros I-X)*
- 15) *...e o que sobretudo me dá maior alegria e confiança, hé ver que em breve tempo me posso encontrar com elle na gloria de Amida: e como eu estou prenhe, não duvido haver em mim algum grave peccado no discurso de minha vida, por onde pela ventura **teria merecido** este genero de morte (Frois, Historia do Japam 2)*

## 5.5. Século XVII

No *subcorpus* do século XVII, encontrámos 938 ocorrências de P-*tinha feito*. Nos 100 casos analisados, este paradigma exprime só acontecimentos ocorridos anteriormente ao ponto de referência, que se encontra no passado, representado ele pelo tempo do relato. (ex. 16).

- 16) *E tornando elle de nouo a me contar todo o successo de seus trabalhos, me relatou todo o discurso de sua vida, & de tudo o mais que **tinha passado**, desde que partira deste reyno até então... (Fernão Mendes Pinto, Peregrinação)*

No que diz respeito a P-*tere* *feito*, no *subcorpus* respetivo, apareceram 50 ocorrências. É curioso notar que, em 44 casos, o paradigma tenha o verbo auxiliar *haver* e apenas em 6 casos o tempo composto seja formado pelo auxiliar *ter*. De acordo com uma pesquisa realizada anteriormente (Hricsina 2017, p. 182), no século XVII, a frequência do verbo auxiliar *haver* deve ter sido mínima. Na maioria dos casos, este paradigma denota situações posteriores ao momento da fala. Visto que muitas vezes se encontra regido por sintagmas verbais como *não duvidar*, *ficar seguro*, *supor*, achamos que os acontecimentos eram considerados como certos ou quase certos (ex. 17 e 18).

- 17) *As minhas desgraças sao de sorte, que ainda àqueles que nao tem em mi tanta parte, como vós tendes no meu sangue e no meu coração, alcançam e molestam. Sendo isto assi, nao posso duvidar de que vos **haverá doído** a minha desesperação; porque isto já nao tem outro nome. Aquele homem me teve em venda como escravo. Todos me tratam como a desfavorecido; e em meus sucessos se tem visto, por mais que eu me cale. (Francisco Manuel de Melo, Cartas familiares)*
- 18) *E por que Vossa Senhoria tenha inteiro conhecimento dos têrmos do negócio, em que fundam que se pedir (e nas boas petições os bons despachos) envio a Vossa Senhoria êsse papel: cópia de outro que a Ene hei oferecido; ficando seguríssimo de que Vossa Senhoria **haverá tomado** por ensaio desta mercê, que lhe peço, as que de antes me tem feito. (Francisco Manuel de Melo, Cartas)*

É curioso também ver que, em alguns casos, após alguns verbos (*supor*, *crer*) nos tenhamos deparado com a elipse da conjunção *que* (ex. 19).

- 19) *Nao há muitos dias que por um framengo, natural de Anveres, que aqui assis-tiu e se foi por via dêsses Estados (seu nome Lucas Vuosterman), escrevi a V. S. ua carta, que ele me prometeu pôr em maos de V. S. e creio o **haverá feito**, se chegou a salvamento. (Francisco Manuel de Melo, Cartas familiares)*

Apenas em 3 casos, o paradigma parece ter denotado acontecimen-tos modalmente considerados prováveis e temporalmente anteriores ao momento da fala (ex. 20).

- 20) *As ordens que de cá foram já **terão chegado**, porque as mais antigas partiram em 10 de Março, e sempre se foram continuando cada vez mais apertadas; entendo que bastarão e **haverão bastado** para segurança dos interessados de fora, e para que os de dentro não tenham perigo da execução. (Padre António Vieira, Cartas)*

Quanto a *P-teria feito*, encontrámos apenas 8 ocorrências no *subcorpus* respetivo. Este paradigma exprime sempre processos incertos ou eventuais que ocorreram anteriormente ao ponto de referência, que se encontra no passado (ex. 21 e 22). A seleção dos verbos auxiliares é semelhante à constatada em *P-terei feito* (5 casos com *haver* e apenas 3 com *ter*).

- 21) *Ontem chegou o correio e hoje parte; a demasiada tardança fazia suspei-tar que o **teriam desvalijado** alguns salteadores franceses, como fizeram estes dias a outro dessa corte que ia para Alemanha; mas eu, depois que li a*



*carta de V. Ex.<sup>a</sup>, entendi que tardou porque me trazia tão boas novas. (Padre António Vieira, Cartas)*

- 22) *...porque os Estrangeiros, quando vissem os Vassallos de Espanha obedientes, não irião ler os acordos de seu arrependimento: sendo certo, q para cessarem as esperanças, e designios que em sua quietação **haverião fundado**, bastava saberse que elles voluntariamente se someterião, ao jugo da vontade real. (Francisco Manuel de Melo, Epanaphora politica primeira)*

## 5.6. Século XVIII

No *subcorpus* do século XVIII, encontrámos 606 ocorrências de *P-tinha feito*. Nos 100 casos analisados, este paradigma denota acontecimentos anteriores ao ponto de referência, que se encontra no passado (ex. 23).

- 23) *Assim se cumprio a profecia em que Oseas **tinha dito** que De Egypto chamaría Deos a seu Filho. (Antonio de Sousa de Macedo, Eva e Ave ou Maria Triunfante)*

Quanto a *P-terei feito*, encontrámos 37 ocorrências. Diferentemente do constatado para o século anterior, em todos os casos o verbo auxiliar está representado pelo verbo *ter*. Tal como acontece nesse século anterior, encontrámos uma elipse da conjunção *que* após os verbos *supor* e *crer* (ex. 24). Este paradigma expressava sempre acontecimentos anteriores (com um valor modal forte) ao ponto de referência, que fica no futuro (ex. 24 e 25).

- 24) *Ao mesmo tempo que aquele me ficou em Roma, me chegou pela nau que veio de Pernambuco antes da frota, outro irmão Jesuíta, mandado a negócio da sua Província, que creio **terá acabado** a tempo de poder-se recolher com a frota do Rio de Janeiro, onde a sua presença me é sumamente necessária para atender a alguns interesses de muita consequência que naquelas partes me sobrevieram, por motivo de casamento. (Alexandre de Gusmão, Cartas)*
- 25) *El-Rei Nosso Senhor, que tem dois Anjos da guarda para acertar em semelhantes eleições, **terá escolhido** a esta hora o sujeito que for mais capaz para os expedientes do seu real serviço. (J. Cunha Brochado, Cartas)*

No que diz respeito a *P-teria feito*, encontrámos 24 ocorrências. Apenas em 4 casos, o verbo auxiliar é *haver*. Além da expressão de acontecimentos incertos ou eventuais que ocorreram anteriormente ao ponto de referência no passado (ex. 26 e 27), nesse século, este paradigma começou a aparecer

em frases condicionais contrafactuais (ex. 28 e 29). No entanto, a sua frequência neste contexto é baixa (5 casos).

- 26) *Se tanto se celebrava a representação, quanto mais se **haveria celebrado** o mesmo dia 9.* (Antonio de Sousa de Macedo, *Eva e Ave ou Maria Triunfante*)
- 27) *Recebo neste correio uma carta de Vossa Mercê em que achei uma novidade que nunca **teria esperado**, por muito que viva persuadido da generosidade e bizzarria de Vossa Mercê.* (Alexandre de Gusmão, *Cartas*)
- 28) *Quanto merecimento **teríeis adquirido** para o Dia do Juízo, se o vosso ouro servisse de fortalecer a castidade vacilante, e não de arruinar a constância da castidade!* (João Baptista de Castro, *A aflição confortada*)
- 29) *Se o ouro nao tivesse corrido tanto da América para a Europa e da Europa para a \*sia, já hoje **teria inundado** a Europa, e **se teria vilipendiado** pela sua abundancia; ele **se teria já feito** de menos preço que o ferro, e **teria perdido** até a mesma qualidade de representativo;* (J. J. da Cunha Azeredo Coutinho, *Obras econômicas*)

## 5.7. Século XIX

No *subcorpus* respetivo, encontrámos 4 007 ocorrências de *P-tinha feito* tanto no PE<sup>18</sup> quanto no PB.<sup>19</sup> Analisámos 100 casos provenientes do PE. Como nos séculos anteriores, em todas as frases analisadas este paradigma expressa situações que ocorreram anteriormente ao ponto de referência, que se encontra no passado (ex. 30).

- 30) *Onde foi? Juliana encolheu os ombros com um sorrisinho. Luísa percebeu. **Tinha ido** a algum amante, a algum amor.* (Eça de Queirós, *O Primo Basílio*)

No que diz respeito a *P-terei feito*, encontrámos 125 ocorrências na totalidade (tanto do PE quanto do PB). Note-se que os casos provenientes do PB são muito mais frequentes que os do PE (102-PB/23-PE). A função predominante deste paradigma é expressão de acontecimentos incertos ou prováveis ocorridos anteriormente ao momento da fala (ex. 31 e 32). Encontrámos apenas 6 casos em que *P-terei feito* denota situações reais ocorridas anteriormente ao ponto de referência, que se encontra no futuro (ex. 33).

18 PE significa Português Europeu e PB Português do Brasil.

19 No corpus [www.corpusdoportugues.org](http://www.corpusdoportugues.org) não é possível separar as ocorrências do PE das do PB do século XIX.

- 31) *O monge, o cavaleiro e todos os habitantes dos paços de Guimarães haviam-se completa e profundamente esquecido do truão, como porventura **terá acontecido** a mais de um dos nossos leitores.* (Alexandre Herculano, *O Bobo*)
- 32) *Ainda mesmo que o pequeno encontrado fosse o teu filho, há que anos **terá morrido** o homem que o encontrou no Tâmega?* (Camilo Castelo Branco, *Maria Moisés*)
- 33) *De qualquer lado que tenha de se decidir a vitória, será disputada, até ao último instante, pelo contendor vencido; a pausa **terá sido** inevitável; a reacção, enérgica; e a crise, violenta.* (Júlio Dinis, *As Pupilas do Senhor Reitor*)

Quanto a P-*teria feito*, encontrámos 428 ocorrências, na totalidade (em PE e PB). Registámos apenas três casos em que o tempo composto é formado pelo verbo auxiliar *haver*. Como no caso de P-*terei feito*, também neste paradigma as ocorrências provenientes do PB são muito mais frequentes do que as do PE (383-PB/45-PE). Nos 45 casos analisados (do PE), encontrámos duas funções deste paradigma: 1. expressão de acontecimentos incertos ou prováveis ocorridos anteriormente ao ponto de referência, que se encontra no passado – 33 ocorrências (ex. 34 e 35); 2. expressão de contrafactualidade (em frases condicionais) – 12 casos (ex. 36 e 37).

- 34) *Veio ao quarto, viu o roupão de Luísa arremessado, achapeleira tombada. Onde **teria ido**? Queixar-se à polícia? Procurar o marido? Cos diabos! Fora estúpida, com o génio!* (Eça de Queirós, *O Primo Basílio*)
- 35) *São assim todos, antes e depois de comer.. na Ajuda. Esta **haveria sido** talvez a única palavra onde o sarcasta revelasse a convicção da sua esmagadora superioridade moral sobre os demais.* (Fialho de Almeida, *Gatos1*)
- 36) *Pois eu sou casada, bem no sabes, senão **teria casado** contigo.* (Camilo Castelo Branco, *A Queda dum Anjo*)
- 37) *Pôs-se a pensar o que **teria sucedido** se tivesse casado com oprimo Basílio.* (Eça de Queirós, *O Primo Basílio*)

## 5.8. Século XX

### Língua escrita

No *subcorpus* da língua escrita do século XX<sup>20</sup>, encontrámos 4 282 ocorrências de P-*tinha feito*, na totalidade (no PE e PB). Nos 100 casos analisados

20 O subcorpus da língua escrita do século XX compõe-se de obras de ficção de autores portugueses e brasileiros do século XX.

(todos provenientes do PE), este paradigma expressa situações anteriores ao ponto de referência que se encontra no passado (ex. 38).

- 38) *Eu fui até Lisboa.. Mas apareci de surpresa lá em casa, vi.. o que nunca **tinha visto**, e fugi para cá, no primeiro comboio. (Francisco Costa, Cárcere Invisível)*

No que diz respeito a *P-terei feito*, no *corpus* respetivo encontramos 71 ocorrências na totalidade (PE e PB). Analisámos 43 casos provenientes do PE, concluindo que em 36 ocorrências o paradigma denota situações prováveis e temporalmente anteriores ao ponto da fala (ex. 39 e 40). Em apenas 7 casos, *P-terei feito* exprime acontecimentos anteriores ao ponto de referência, que fica no futuro (ex. 41).

- 39) *Calou-se o velho? Alguém **terá sido** acusado em meu lugar? Ainda hoje o não sei e nunca o saberei, provavelmente. (Francisco Costa, Cárcere Invisível)*
- 40) *Que pena ela me faz! Que irá na sua vida? Que se **terá passado** ontem? Até que ponto serão exactas as nossas apreensões? (Fialho D'Almeida, A Cidade do Vício)*
- 41) *O revisor é observador bastante competente e sensível para, num simples relance do olhar, recolher uma informação tão completa, podemos mesmo admitir a hipótese de que algum dia **terá encontrado** no espelho da sua casa uns olhos assim, os seus próprios, não seria preciso dizê-lo, porém não vale a pena perguntar-lho, que, dele, o que mais nos interessa é o presente, e, se do passado uma lembrança, muito menos o seu do que, do passado geral, a parte modificada pela palavra impertinente. (José Saramago, História do Cerco de Lisboa)*

Quanto a *P-teria feito*, encontramos 524 ocorrências na totalidade (nos PE e PB), sendo 217 casos provenientes do PE. Nos 100 casos analisados (do PE), registámos 76 exemplos em que este paradigma denota acontecimentos prováveis e temporalmente anteriores ao ponto de referência passado (ex. 42 e 43). Em 24 casos, *P-teria feito* aparece com o valor condicional contrafactual (ex. 44).

- 42) *Passaram minutos, Néné não se mostrava, era como se Néné se houvesse afundado pela terra abaixo, carago, mas isto é demais, ela já **teria ido** para a escola? (Aquilino Ribeiro, A Via sinuosa)*
- 43) *Noitadas misteriosas, não é exagero nenhum, visto que a casa estava desabitada. Meses antes, Sandra Lulu tinha falado de um velho que **teria visto** na escada a marinhar pelo corrimão e de passos arrastados a seguir no andar de*

*cima: daquela casa foi o único sinal de vida que ela sentiu desde há muitos meses a esta parte. (José Cardoso Pires, A república dos corvos)*

- 44) *Se o Dr. Quaresma tivesse dito qualquer coisa, eu **teria respondido** qualquer coisa; teria tido a que adaptar a minha razão e a minha voz. (Fernando Pessoa, O Roubo da quinta das vinhas)*

## Língua falada

No *subcorpus* da língua falada do século XX, encontramos 552 ocorrências de *P-tinha feito* na totalidade (em duas variantes do Português). Nos 100 casos analisados, este paradigma exprime apenas acontecimentos anteriores ao ponto de referência, que se encontra no passado (ex. 45). Ou seja, a tendência à expansão funcional deste paradigma a frases condicionais contrafactuais não foi confirmada.

- 45) *A última pessoa a estudar as algas tinha sido o Professor Mesquita Rodrigues, que nessa altura já não estava no Instituto Botânico de Coimbra, uma vez que **tinha ido** dirigir o Laboratório de Botânica da Universidade de Lourenço Marques, em Moçambique. (Jorge Rino – entrevista)*

No *subcorpus* respetivo, encontramos apenas um total de 32 ocorrências de *P-terei feito* (nos PE e PB). Analisámos 31 casos provenientes do PE (no PB, foi registada apenas uma ocorrência), constatando que este paradigma denota apenas situações prováveis anteriores ao ponto da fala (ex. 46 e 47). Não encontramos, assim, nenhum exemplo em que o paradigma em questão exprimisse uma situação anterior a um ponto de referência futuro.

- 46) *Felizmente, nessa altura, a Universidade de Aveiro tinha como Reitor o Prof. Júlio Pedrosa, uma pessoa com uma visão muito clara daquilo que queria para a UA. Eu creio que ele **terá sido** uma das pessoas na Universidade de Aveiro que compreendeu que o tipo de Biologia que se fazia no fim do século XX ultrapassava largamente aquilo que se fazia no Departamento de Biologia da Universidade de Aveiro. (Edgar Figueiredo da Cruz e Silva – entrevista)*
- 47) *Segundo o que foi publicado, você **terá dito** que o Barcelona tratava os jogadores como mercadoria.. (Vitor Baía – entrevista)*

No que diz respeito a *P-teria feito*, encontramos 123 ocorrências, na totalidade (nos PE e PB), sendo 51 casos provenientes do PE. A análise destes exemplos mostrou-nos que em 27 casos este paradigma tem um sentido

condicional (ex. 48 e 49) e, em 24 casos, exprime acontecimentos prováveis ocorridos anteriormente ao ponto de referência, que se encontra no passado (ex. 50 e 51). A proporção existente entre as duas funções é, assim, bastante equilibrada.

- 48) *Se fosse no seu tempo de ministro dos Negócios Estrangeiros este caso **teria sido** conduzido de que maneira?* (Durão Barroso - 97-08-11-43 – entrevista)
- 49) *É verdade, nisso errei e não o repetirei. Se o tivesse feito mais próximo, eventualmente ninguém **teria acalentado** expectativas e não **teria ocorrido** alguma complicação interna como sucedeu.* (Manuela Teixeira - 97-01-08-51 – entrevista)
- 50) *José Barata-Moura constou que ponderou a possibilidade de não se recandidatar a um segundo mandato. O que o **teria levado** a considerar essa possibilidade e o que o levou a apresentar nova candidatura?* (Maia Nogueira – entrevista)
- 51) *O que é que aconteceu realmente no caso relatado pela imprensa em que o senhor **teria atribuído**, enquanto presidente do CRSS de Faro, um subsídio a uma fundação dirigida por si próprio?* (Carlos Andrade - 97-03-20 – entrevista)

## 6. A frequência e o emprego dos tempos compostos na evolução da língua portuguesa

Nesta parte, vamos resumir a frequência absoluta<sup>21</sup> e o emprego de *P-terei feito* e *P-teria feito* na história do Português Europeu, mostrando os resultados da análise do *corpus* em gráficos.

O gráfico 1 mostra a evolução da frequência dos valores modo-temporais observados em *P-terei feito* na história do Português Europeu (o valor do futuro é representado a cinzento, expressão de situações ocorridas antes das outras futuras a cor de laranja, o valor de probabilidade no passado a azul e a expressão de situações ocorridas depois das outras futuras a amarelo).

---

21 Visto que o tamanho do subcorpus é diferente para cada século, os dados não mostram uma proporção de frequência de ocorrências entre vários séculos de uma maneira exata.

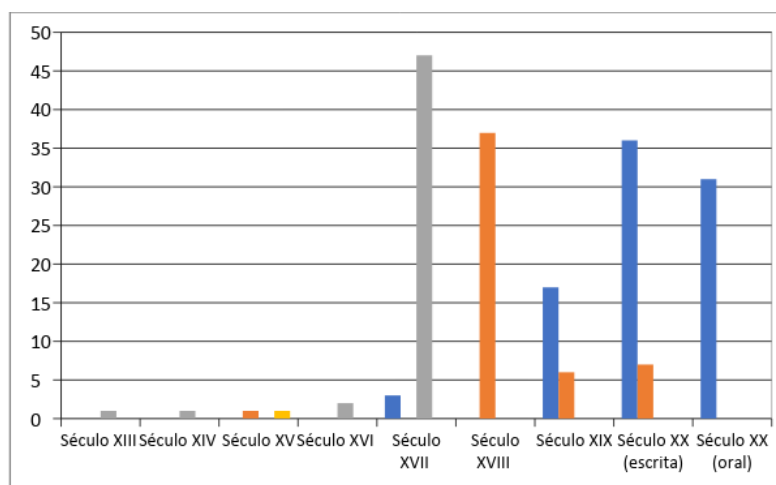


Gráfico 1. Valores modo-temporais de P-terei feito na história da língua portuguesa

O gráfico 2 mostra a evolução da frequência dos valores modo-temporais observados em *P-teria feito* na história do Português Europeu (o mais-que-perfeito com valor de probabilidade é representado a azul e o do condicional a cor de laranja).

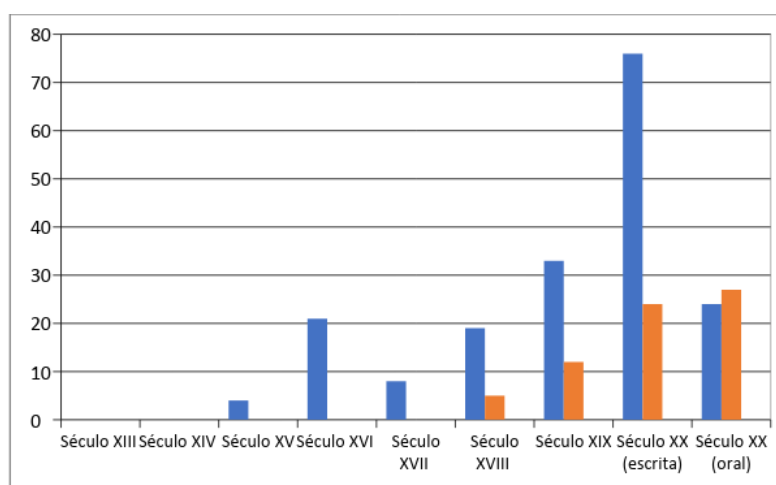


Gráfico 2. Valores modo-temporais de P-teria feito na história da língua portuguesa

## 7. Conclusões

A pesquisa efetuada no *corpus* linguístico [www.corpusdoportugues.org](http://www.corpusdoportugues.org) traz os seguintes resultados:

1. *P-tinha feito* existe na língua portuguesa desde o século XIII, mesmo não sendo muito frequente neste século. Em toda a evolução do Português, tem tido uma única função: a expressão de acontecimentos anteriores a um ponto de referência que se encontra no passado.
2. Nos séculos XIII a XVI, no *corpus*, *P-terei feito* aparece de uma maneira esporádica e exprime sempre acontecimentos posteriores ao tempo da fala. Só no século XVII começa a aparecer mais frequentemente, continuando a denotar situações futuras. No século seguinte, o paradigma muda de conteúdo modo-temporal e começa a exprimir situações anteriores a um ponto de referência que se encontra no futuro. No século XIX, adquire uma nova função, denotando situações prováveis ou incertas ocorridas anteriormente ao tempo da fala, enquanto a frequência da expressão dos factos ocorridos anteriormente ao ponto de referência futuro vai diminuindo. Esta tendência é ainda mais forte no século XX. Na língua falada não encontramos nenhuma frase em que este paradigma denotasse um acontecimento ocorrido antes do ponto de referência futuro.
3. *P-teria feito* tem a frequência mais baixa dos paradigmas analisados. No *corpus*, verifica-se que começa a aparecer só no século XV e de forma esporádica. Até ao século XVIII, altura em que a sua frequência começa a crescer, denota sempre ações eventuais ocorridas antes de um ponto de referência que se encontra no passado. No século XVIII, começa a aparecer também em frases condicionais contrafactuais, apesar de a sua frequência neste contexto ser muito baixa. Nos séculos seguintes, a frequência do sentido condicional deste paradigma vai crescendo.

## Referências

- Berta, T. (2016). Contribuição para a história da eliminação da concordância do participio nos tempos compostos do Português. In J. C. Ramos, Š. Grauová & J. Jindrová (Eds.), *Língua portuguesa na Europa central: Estudos e perspetivas* (pp. 174–183). Praga: Karolinum.
- Brocardo, M. T. (2014). *Tópicos de História da Língua Portuguesa*. Lisboa: Edições Colibri.
- Cunha, C. & Cintra, L. F. L. (1999). *Nova Gramática do Português Contemporâneo*. (10ª ed.) Lisboa, Portugal: João Sá da Costa.



- Dias, A. E. da S. (1933). *Syntaxe Historica Portuguesa*. (2ª ed.) Lisboa: Livraria Clássica Editora.
- Duarte, I. M. (2009). *Futuro perfeito e condicional composto: Mediativo no discurso jornalístico em português europeu e em português brasileiro*. Disponível em: <<https://repositorio-aberto.up.pt/bitstream/10216/13504/2/congressoabralinisabelduarte000071277.pdf>>. Consultado em: 15 jan. 2018.
- Hricsina, J. (2017). Evolução do verbo auxiliar no Português Europeu. *Études romanes de Brno*, 38(1), 165–184.
- Huber, J. (2006). *Gramática do Português Antigo*. (2ª ed.) Lisboa: Fundação Calouste Gulbenkian.
- Mattos e Silva, R. V. (2008). *O português Arcaico-Volume I – Léxico e morfologia*. Lisboa: Imprensa Nacional-Casa da Moeda.
- Oliveira, F. (2004) Tempo e aspecto. In M. Mira & M. Helena (Eds.), *Gramática da Língua Portuguesa* (pp. 127–178). Lisboa: Caminho.
- Oliveira, F. (2013). Tempo verbal. In Raposo, Eduardo Paiva (Eds.), *Gramática do Português – Volume I* (pp. 509–553). Lisboa: Fundação Calouste Gulbenkian.
- Paiva Raposo, E. B. (2013). Verbos auxiliares. In E. P. Raposo (Ed.), *Gramática do Português – Volume II*. (pp. 1221–1284). Lisboa: Fundação Calouste Gulbenkian.
- Ribeiro, I. (1996). A formação dos tempos compostos: a evolução histórica das formas ter, haver e ser. In I. Roberts & M. Kato (Eds.), *Português brasileiro: uma viagem diacrônica: Homenagem a Fernando Tarallo* (pp. 343–386). Campinas: Editora da Unicamp.
- Svobodová, I. (2014). *Morfologie současného portugalského jazyka II. Sloveso*. Brno: Masarykova univerzita.
- Tláškal, J. (1978). Remarques sur le futur en portugais contemporain. *Philologica pragensia*, 21(4), 204–213.
- Zavadil, B., Čermák, P. (2010). *Mluvnice současné španělštiny*. Praga: Karolinum.

[recebido em 1 de abril de 2018 e aceite para publicação em 8 de janeiro de 2019]

# ANÁLISE CONTRASTIVA DAS FORMAS DE TRATAMENTO AO INTERLOCUTOR NO TEATRO BRASILEIRO E PORTUGUÊS DOS SÉCULOS XIX E XX

## CONTRASTIVE ANALYSIS OF THE FORMS OF ADDRESS IN THE BRAZILIAN AND PORTUGUESE THEATER OF THE 19<sup>TH</sup> AND 20<sup>TH</sup> CENTURIES

Ana Carolina Morito Machado\*  
ana.machado@ifrj.edu.br

Este trabalho tem como objetivo analisar o comportamento das formas de tratamento ao interlocutor, que, nesta pesquisa, são consideradas *Tradições Discursivas*, no Português brasileiro e europeu dos séculos XIX e XX. Para tal, investiga-se a distribuição dessas estratégias, à luz da Sociolinguística Variacionista, em vinte e nove peças teatrais escritas ao longo desse período. Para auxiliar na compreensão de questões relativas às relações sociais e a motivações pragmáticas, foram utilizadas questões discutidas pelo trabalho *The pronouns of power and solidarity*, de Brown e Gilman (2003 [1960]). Os resultados apontam para uma similaridade entre o sistema de tratamento no Português brasileiro (PB) e do Português europeu no terceiro quarto do século XIX, e para um progressivo distanciamento entre as duas variedades, a partir dos últimos anos do século XIX.

**Palavras-chave:** Estratégias de referência à segunda pessoa. Tradições Discursivas. Sociolinguística Variacionista. Cortesia.

The main goal of this work is to analyze the behavior of address forms, which, in this research, are considered Discursive Traditions in Brazilian and European Portuguese of the nineteenth and twentieth century. To this purpose, it was investigated the distribution of those address forms, using the Variational Sociolinguistics, in twenty nine plays written in those centuries. To better understand the questions related to social relations and pragmatic motivations, it was also used the paper. The pronouns of power and solidarity, of Brown e Gilman (2003 [1960]). The results have shown a similarity between the systems of address forms in the Brazilian Portuguese and the European Portuguese in the third quarter of nineteenth century.

---

\* Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro, Brasil.

However, they point out a continuous distancing between the two Portuguese forms in the last years of nineteenth century.

**Keywords:** Reference strategy of second person. Discursive Traditions. Variational Sociolinguistics. Courtesy.



## 1. Introdução

Este estudo tem como objetivo contribuir, juntamente a outros trabalhos sobre o tema, para a descrição do sistema de tratamento ao interlocutor no Português brasileiro (doravante PB) e no Português europeu (doravante PE) dos séculos XIX e XX. Sabe-se como diversos estudos sobre questão do tratamento mostraram, que, para entender o percurso das estratégias de referência ao interlocutor em uma língua, deve-se investigar a estreita relação que há entre evolução sócio-histórica, mudança linguística e tradição textual, pois os textos têm história própria, independente da história da língua (mais adiante será tratada essa questão).

Segundo Goffman (2007 [1959]), todo homem está a todo momento representando um papel, mesmo que de modo não consciente. É através dos papéis que um indivíduo desempenha que os outros o conhecem e que ele conhece a si próprio. De certo modo, o papel que esse indivíduo se esforça para representar é o que ele é de fato ou, ao menos, o que ele deseja ser. Os papéis, na perspectiva de Robinson (1977 [1972], p. 114), correspondem “ao conjunto de comportamentos prescritos para (ou esperáveis de) uma pessoa que ocupe certa posição na estrutura social”. Sendo assim, o papel é um constructo institucionalizado em uma determinada cultura. Nesse constructo, a linguagem verbal desempenha uma função essencial na atribuição dos papéis sociais, pois, dentre as diversas formas linguísticas que contribuem para a reivindicação de um papel por parte do falante e a concessão pelo falante de um papel a seu interlocutor, encontram-se as formas de tratamento. Em outras palavras, a escolha por uma forma de tratamento está associada aos papéis sociais desempenhados pelos participantes que se encontram inseridos em uma situação interacional.

Além dessa intrínseca relação entre formas de tratamento e papéis sociais, acredita-se, também, que há uma forte ligação entre essas formas e

movimentos de conservação e inovação na língua. A fim de entender melhor a questão, parte-se do pressuposto de que as línguas são compostas por Tradições Discursivas (doravante TDs), que são formas textuais evocadas em um contexto específico e que, pela repetição desses elementos linguísticos, nesse mesmo contexto, ao longo da história, tornam-se “cristalizadas” em uma determinada situação comunicativa. As TDs estão presentes na língua de um modo geral. Um dos domínios em que ocorrem TDs é no paradigma das formas de tratamento, uma vez que o emprego dessas formas pressupõe repetição de fórmulas linguísticas que são evocadas em situações de interação entre papéis sociais específicos. Entretanto, tem-se em vista também que tanto as formas de tratamento ao interlocutor quanto às relações sociais mudam ao longo do tempo, o que acarreta mudanças na língua neste domínio.

Tomando como base essas questões, este trabalho tem como objetivo observar, em peças teatrais, o tratamento ao interlocutor na variedade do Rio de Janeiro e de Lisboa dos séculos XIX e XX. Pretende, sobretudo, analisar, no teatro, como fórmulas linguísticas consideradas inaceitáveis para determinados contextos comunicativos transformam-se em recorrentes, tornando-se, assim, índices de “flexibilização” das relações interpessoais.

A fim de alcançar esse propósito, serão analisadas as estratégias de referência ao interlocutor em duas amostras de textos dramáticos, ambientados no Rio de Janeiro e em Lisboa, à luz do modelo das Tradições Discursivas (TDs), da Teoria da Variação (Weinreich, Labov & Herzog 1968) e da Teoria do Poder e Solidariedade (Brown & Gilman 1960).

## **2. Pressupostos teóricos**

### **2.1. A união entre o modelo das Tradições Discursivas e a teoria Sociolinguística: breves considerações**

Tendo em vista que o objetivo principal deste estudo é descrever as alterações operadas no sistema de tratamento ao interlocutor no PB e no PE dos séculos XIX e XX, fez-se a opção pela união entre dois modelos teóricos que discutem a mudança linguística: o das Tradições Discursivas e da Sociolinguística de base laboviana. É de fundamental importância ressaltar que não serão apresentados profundamente os dois modelos teóricos, mas somente o que, neste trabalho, é importante para a explicação das mudanças operadas no paradigma das formas de referência ao interlocutor.

Antes de apresentar as Tradições Discursivas, parece válido demonstrar um exemplo dado por Kabatek (2006) para explicitar esse conceito. Para o

autor, em um encontro pela manhã entre duas pessoas conhecidas em que há um desejo claro de se saudarem, não basta somente encontrar uma expressão da língua portuguesa; é necessário que se produza um enunciado como “bom dia”, seguindo, assim, uma tradição que é estabelecida pela cultura e que se encontra além das regras gramaticais e do simples conhecimento dos itens lexicais. Esse emprego de “bom dia” seria uma TD, pois apresenta os dois elementos fundamentais para o estabelecimento de uma TD: a *repetição*, que ocorre quando um texto estabelece uma relação com outros textos em um determinado momento da história; e a *evocação*, que se dá com a repetição dos conteúdos temáticos que são tratados nos textos. Ainda de acordo com o autor, um texto historicamente situado se relaciona com diversos elementos de seu ‘contexto’. Esse ‘contexto’ apresenta inegavelmente um forte conteúdo semântico e pode adquirir um valor simbólico. A repetição (quase sempre parcial) dos elementos contextuais da primeira enunciação evoca a repetição do texto (ou, ao menos, traz a lembrança do primeiro texto, ou mais amplamente, da TD).

Assim sendo, segundo Koch e Oesterreicher (1997 *apud* Kabatek 2006), *Tradições Discursivas* são todas as formas e fórmulas comunicativas que são recorrentes, tradicionais, cujas fronteiras estão além das estabelecidas para as línguas históricas. Para os autores, ao praticar a atividade do falar, o indivíduo se submete a dois filtros concomitantes até transformar o que deseja em um enunciado, um correspondente à língua histórica e outro, às TDs.

Diante do exposto, considera-se que as estratégias de referência à segunda pessoa são TDs, pois evocam um uso: se o interlocutor é, por exemplo, uma pessoa da idade semelhante a do falante e se tem com ela certo tipo de intimidade, a forma de tratamento mais indicada, evocada, no Português brasileiro contemporâneo, é *você*, e, embora, não haja nenhum impedimento gramatical ou lexical para que utilize uma estratégia como *o(a) senhor(a)*, a tradição recomenda que não se empregue essa forma.

Dessa maneira, a associação das noções de estratégias de referência ao interlocutor ao conceito de TDs pode ser explicada na medida em que se sabe **(a)** que, aos falantes de uma determinada língua, estão disponíveis inúmeras formas de natureza nominal, pronominal ou, até mesmo, verbal, de se dirigir à segunda pessoa do discurso, e **(b)** que, entretanto, não é difícil para qualquer falante de uma língua distinguir, por exemplo, quais dessas formas são mais adequadas ao domínio da formalidade ou da informalidade, qual estratégia deve ser utilizada em um tipo de relação que envolva poder ou não, que seja institucional ou não.

Como já visto, está bastante evidente que a utilização de certas estratégias de se dirigir ao interlocutor está intimamente ligada às relações sociais

que se desenvolvem culturalmente e que, dessa maneira, ao se modificarem, acarretam alterações na língua. Quando ocorrem mudanças na história externa de uma língua, inauguram-se novas necessidades comunicativas, que motivam o surgimento de novas TDs. Essas novas TDs provocam modificações no âmbito da língua e, desse modo, a ligação entre a história externa e interna da língua está nas TDs.

A partir dessa relação entre história externa e história interna, neste trabalho, acredita-se que a mudança no domínio das formas de tratamento possa ocorrer de três modos distintos. O primeiro modo seria a *inserção*, que ocorre quando há o aparecimento de uma ‘nova’ situação comunicativa, que necessita ser marcada por uma forma de tratamento específica. Já o segundo modo corresponde à *expansão linguística*, que se relaciona ao alargamento do uso de uma forma linguística para novos contextos situacionais (que, inicialmente, têm algo em comum). Por fim, o terceiro modo é aquele em que há *mudança semântica* da forma de tratamento. Nesse modo, uma forma expande-se para contextos comunicativos que possuem uma natureza distinta do contexto situacional a que a forma estava associada anteriormente. A natureza desse novo contexto pode inaugurar, então, uma nova semântica para a forma comunicativa. Utilizando o exemplo da forma *Vossa mercê*, é possível notar que, paulatinamente, ao se expandir e reduzir sua substância fonológica, essa forma atinge novos contextos situacionais, como a relação entre iguais não íntimos, adquirindo um novo significado relacionado, agora, à noção de simetria.

Os dois últimos modelos — o da expansão e o da mudança da semântica — parecem acarretar uma variação linguística, uma vez que, ao ser evocada em uma nova situação comunicativa e modificar sua essência, a ‘nova forma’ passa a representar um tipo específico de relação que não deixa de evocar também a ‘antiga forma’. Para entender melhor a questão da variação e da mudança linguística, utilizaram-se alguns pressupostos da Sociolinguística de base laboviana.

Na obra clássica *Fundamentos empíricos para uma teoria da mudança linguística*, de 1968, Weinreich, Labov e Herzog afirmam que as escolhas realizadas pelos indivíduos ou por um conjunto deles não são aleatórias, mas condicionadas por princípios tanto internos à estrutura linguística quanto inerentes ao sistema social desta. Além disso, os teóricos dessa corrente consideram que, embora possa haver variação no plano da fala (no sentido de ser uma escolha individual), a variação linguística, em geral, situa-se no plano do sistema.

Seria justamente a variabilidade inerente ao sistema o fato que explicaria as mudanças na língua no espaço temporal, sem perda da estruturalidade,

visto que, na maioria das vezes, os falantes não percebem que estão vivenciando tais alterações. Para essa sensação de imutabilidade, é essencial a gradualidade das mudanças linguísticas. A esta ideia estão intimamente ligados os processos de coexistência e concorrência entre formas ‘novas’ e ‘antigas’ na língua, em que há uma fase de transição em que formas ‘novas’ e ‘antigas’ coexistem, estão em variação, e pode ser que, em um dado momento, a forma mais ‘nova’ suplante a ‘antiga’ e essa forma ‘antiga’ venha a desaparecer. Foi o que, aparentemente, ocorreu com *você* e *tu* na variedade do Rio de Janeiro do PB durante boa parte do século XX. As formas coexistiram no século XIX e nas duas primeiras décadas do século XX, com o aumento progressivo dos índices de *você* ao longo desse tempo, no entanto, por volta dos anos de 1930, *você* suplantou *tu* e este praticamente desapareceu até a década de 1990.

## 2.2. A teoria do Poder e Solidariedade: breves considerações

Segundo Brown e Gilman (2003[1960]), o par conceitual *poder* e *solidariedade* está presente em todas as formas de interação verbal. O *poder* pode ser compreendido como o controle que uma pessoa exerce sobre outra em uma determinada situação interativa. Desse modo, para que haja uma relação de poder, é necessário que pelo menos duas pessoas estejam interagindo socialmente e que a relação entre ambas não seja recíproca, simétrica. A necessidade de não reciprocidade da relação se deve ao fato de que todos os participantes da interação não podem ter poder na mesma área de comportamento. Sendo assim, o poder está presente em relações assimétricas, diferenciais ou não recíprocas e essa hierarquia pode ser observada em atributos como idade, geração e autoridade (o pai é superior ao filho, o professor, ao aluno, o patrão, ao empregado).

Ao contrário do que ocorre em uma relação de poder em que o conceito de hierarquia é de fundamental importância para entender a assimetria no tratamento, na *solidariedade*, pressupõe-se a existência de forças iguais, de um mesmo nível de hierarquia social decorrente de relações sociais recíprocas ou simétricas. Essas relações simétricas derivam fundamentalmente dos atributos de sexo, parentesco e filiação de grupo que, por sua vez, estão ligados às ideias de afinidade, semelhança, afeto e agrado.

Há, portanto, dois tipos de relação de poder — uma em que o emissor exerce poder sobre o receptor e outra em que é o receptor que é o detentor do poder frente ao emissor — bem como há duas formas de relação

de solidariedade — uma em que a solidariedade entre os participantes da situação comunicativa se faz presente e outra em que não é possível observar solidariedade.

Segundo os autores, o uso de formas *V* (como *vous*, em francês) está intimamente ligado a relações simétricas em que os componentes da ação não apresentam afinidades (nas relações não solidárias), e a situações assimétricas em que o emissor se encontra em uma situação hierarquicamente inferior à do receptor. Formas *T* (como *tu*, em francês), ao contrário, estão a serviço de relações simétricas recíprocas e solidárias e de situações assimétricas em que o emissor exerce alguma forma de poder sobre o receptor.

Trabalhos como os de Cintra (1986 [1972]), Faraco (1996), por exemplo, descrevem que, nos primeiros séculos de sua trajetória, *você* apresentava uma semântica fortemente ligada às noções de assimetria e cortesia — comportamento semelhante ao das denominadas formas *V*; entretanto, aparentemente, essa forma passou a pertencer a domínios que antes eram exclusivos das formas *T*, como evidenciam trabalhos como o de Paredes Silva (1999).

### 3. O Corpus

Para a análise de estratégias de referência ao interlocutor, é necessária a escolha de um gênero textual que permita a interação direta entre emissor e receptor em uma situação comunicativa. Sabe-se que a modalidade que abriga, por excelência, esse gênero textual é a oral; no entanto, em épocas mais distantes da contemporaneidade, a apreensão dessa modalidade só é possível através da representação desta na escrita. A fim de atender todos esses requisitos, o gênero peça teatral apresenta-se como um dos mais adequados para a análise linguística desse fenômeno, uma vez que, apesar de ser um texto escrito, é destinado à representação — só adquire vida ao se corporificar numa encenação, sua relação com o público, *grosso modo*, não se dá através da leitura, e sim da encenação de atores que agem sobre a composição, inicialmente, escrita em ação dialogada. Sua importância para os estudos linguísticos é inegável, pois, apesar de constituir um texto escrito, fruto da percepção individual de seu criador, busca representar usos linguísticos próprios das relações que se estabelecem no interior da organização social em que seus autores estão inseridos. Por isso, constitui um valioso material de análise linguística.

Por essas razões, com o objetivo de observar as tendências no comportamento das estratégias de referência ao interlocutor no PB e no PE dos



séculos XIX e XX, serão analisadas vinte e nove peças teatrais — quatorze ambientadas no Rio de Janeiro, e quinze em Lisboa neste período. Entre as peças teatrais, busca-se privilegiar as obras que buscam retratar a vida familiar dos cariocas e lisboetas nesses séculos, constituindo-se de cenários que compreendem, preponderantemente, ambientes privados — casa, pensão onde residem os personagens — e apresentando, em geral, relações íntimas, uma vez que devem se ocupar de situações corriqueiras do cotidiano.

A seguir, encontram-se elencadas as peças selecionadas:

**Quadro 1. A constituição da amostra brasileira**

<b>Amostra do português brasileiro</b>		
<i>Peça</i>	<i>Autor</i>	<i>Data</i>
OS CIÚMES DE UM PEDESTRE	Martins Pena	1846
O DEMÔNIO FAMILIAR	José de Alencar	1857
AMOR COM AMOR SE PAGA	França Júnior	1870
O DEFEITO DE FAMÍLIA	França Júnior	1870
NÃO CONSULTE MÉDICO	Machado de Assis	1896
QUEBRANTO	Coelho Neto	1908
O SIMPÁTICO JEREMIAS	Gastão Tojeiro	1918
O HÓSPEDE DO QUARTO Nº 2	Armando Gonzaga	1937
DONA XEPA	Pedro Bloch	1952
TODA DONZELA TEM UM PAI QUE É UMA FERA	Gláucio Gill	1962
O GENRO QUE ERA NORA	Aurimar Rocha	1972
COMUNHÃO DE BENS	Alcione Araújo	198-
INTENSA MAGIA	Maria Adelaide Amaral	1995
SÍNDROMES	Maria Carmen Barbosa e Miguel Falabella	2003

Quadro 2. A constituição da amostra portuguesa

Amostra do português europeu		
<i>Peça</i>	<i>Autor</i>	<i>Data</i>
AS ASTÚCIAS DE ZANGUIZARRA	Ricardo José Fortuna	1819
O BEATO ARDILOSO	José Joaquim Bordalo	1825
<i>SIMILIA SIMILIBUS</i>	Júlio Dinis	1858
O FERRO VELHO	P.C. D'Alcantara Chaves	1866
QUEM DESDENHA...	Manuel Joaquim Pinheiro Chagas	1874
FIM DE PENITÊNCIA	Marcelino António da Silva Mesquita	1895
O TIO PEDRO	Marcelino António da Silva Mesquita	1902
ZILDA	Alfredo Cortez	1921
VIVA DA COSTA	Vasco Mendonça Alves	1925
TRÊS GERAÇÕES	Ramada Curto	1931
A PRIMA TANÇA	Alice Ogando	1934
É URGENTE O AMOR	Luiz Francisco Rebello	1956
O HOMEM DO QUIOSQUE	Tomaz de Figueiredo	1958
OS OUTROS	Jaime Salazar Sampaio	1965
ANTÓNIO, UM RAPAZ DE LISBOA	Jorge Silva Melo	1995

## 4. A análise dos resultados

### 4.1. As formas de tratamento ao interlocutor em análise

Analisaram-se as estratégias de referência ao interlocutor na função de sujeito explícito ou desinencial associadas a verbos nos modos indicativo e subjuntivo. A seguir, há exemplos das estratégias sob investigação.

#### . TU

(01) A AVÓ — Ó Rosa, ***tu*** não sabes o que *estás* a dizer!... Olha que ***tu*** estás ainda em muito boa idade... (PE – Três gerações, 1931)

(02) EMÍLIA — Acredita, Carlos. Não posso levar à paciência que ***tu***, tão trabalhador e amigo da família, vás prender-te a uma criatura inútil, que só cuida de leitura e arrebiques... ***Vais*** ser um desgraçado! (PE - Zilda, 1921)

## . VÓS

(03) PEDESTRE — *Sim, sim, está morta... Mas vós lhe dareis vida por um navio... vinde... silêncio... Dar-vos-ei um navio que ela me fez perder...* (PB - *Os ciúmes de um pedestre*, 1846)

## . VOCÊ

(04) VEIGA — *Salve! Zilda! Ó Zilda! Você, a sério, não gosta destas gravuras?* (PE — *Zilda*, 1921)

(05) MACEDO — *Você não quer admitir que sou um coroa ainda muito enxuto, não é? Mas pode deixar, meu bem. Eu vou morrer apaixonado por você.* (PB - *O genro que era nora*, 1972)

## . O(A) SENHOR(A)

(06) ZEZE — *O senhor vive dizendo que é o que é mas nunca conseguiu aceitar os outros como são.* (PB - *Intensa Magia*, 1995)

(07) RAPARIGA — *Sou aquela rapariga a quem o senhor há bocado ligou o pulso... Está a ver a ligadura?[...]* (PE – *Os outros*, 1958)

## . VOSSA MERCÊ

(08) LUÍSA — *Ora, meu pai, vossa mercê está persuadido que eu havia temerariamente levantado um testemunho ao Tio Lourenço?* (PE – *O beato ardiloso*, 1825)

## . VOSSA EXCELÊNCIA

(09) AZEVEDO — *Aqui passa V. Ex.<sup>a</sup> naturalmente as tardes, conversando com suas flores, em doce e suave rêverie!* (PB - *O demônio familiar*, 1857)

(10) CHEFE — [...] Sr. Dr. Jorge Fonseca... É V. Ex.<sup>a</sup>? Quer ter a bondade de entrar?

[...]

CHEFE — [...] V. Ex.<sup>a</sup> chegou muito a tempo.

[...]

CHEFE — Tenha a bondade... De certo não ignora o pedido que me levou a pedir-lhe que visse aqui... (PE – *É urgente o amor*, 1956)

## . VOSSA SENHORIA

(11) PEDESTRE — [...] *Que ordena vossa senhoria?* (PB – *Os ciúmes de um pedestre*, 1846)

### . SINTAGMAS NOMINAIS

(12) DR. MATEUS — O Sr. Tomás Bento dá licença. (PE - *Similia Similibus*, 1858)

(13) TIBÚRCIO — Mas o doutor disse ao Honorato que eu tinha um filho hospedado aqui. (PB - *O hóspede do quarto número dois*, 1937)

(14) CHEFE — A menina era muito amiga dela, não é verdade? (PE – *É urgente o amor*, 1956)

(15) PEDESTRE — Ah, a menina tem namorados, recebe cartinhas e quer casar-se contra a minha vontade [...] (PB - *Os ciúmes de um pedestre*, 1846)

(16) JOSINO — [...] Vovó deve compreender que um homem, em vésperas de casamento, tem obrigação de retemperar-se para resistir aos encargos da vida conjugal. (PB – *Quebranto*, 1908)

(17) MANUEL — Ora a Zilda vive aqui dentro uma vida a que procura inutilmente aclimar-se... (PE - *Zilda*, 1921)

## 4.2. A distribuição geral das formas

A análise das formas de tratamento ao interlocutor na função de sujeito explícito ou desinencial nas vinte e nove peças ambientadas no Rio de Janeiro e em Lisboa dos séculos XIX e XX, tomando-se apenas os dados relacionados a formas verbais não-imperativas, resultou um total de 7148 dados.

É fundamental ressaltar que, a fim de sintetizar os resultados obtidos na análise das amostras brasileira e portuguesa, agruparam-se os dados aproximadamente por quartos de século. Na amostra brasileira, as obras compreendidas entre os anos de 1846 e 1870 apresentam-se como XIX (3); a de 1896, como XIX (4); as de 1908 e 1918, como XX (1), as de 1937 e 1952, como XX (2), as de 1962 e 1972, como XX (3), e as de 1980, 1995 e 2003, como XX (4). Na amostra portuguesa, as peças de 1819 e 1825 foram agrupadas como XIX (1/2); as de 1858 e 1866, como XIX (3); as 1874 e 1895, como XIX (4); as de 1902, 1921 e 1925, como XX (1); as 1931 e 1934, como XX (2); as de 1956, 1958 e 1965, como XX (3); e a de 1995, como XX(4).

A seguir, encontra-se a distribuição geral dos dados.



Tomando as peças de teatro analisadas, inicialmente, é importante ressaltar a diferença na quantidade de formas encontradas no século XIX e no século XX. Tomando os extremos da tabela, pode-se afirmar, de acordo com os resultados obtidos nas amostras sob análise, que, se na primeira metade do século XIX, há seis estratégias de tratamento ao interlocutor, no último quarto do século XX, existem praticamente apenas três — *tu*, *você* e *o(a) senhor(a)*.

Além dessa ‘perda de diversidade’ das estratégias de tratamento ao interlocutor, quando se analisa o comportamento do pronome *tu*, na amostra brasileira, embora essa forma predomine no terceiro quarto do século XIX, seus índices decrescem no final desse século e nas primeiras décadas dos anos de 1900 até seu uso praticamente desaparecer a partir do segundo quarto do século XX. Já, na amostra portuguesa, observa-se que essa forma prepondera em todas as épocas sob investigação, apresentando índices entre 40% e 60% no século XIX, e sempre superiores a 65% nos períodos do século XX. Dessa maneira, além de sua superioridade, destaca-se o aumento progressivo de seus índices, o que pode estar relacionado à flexibilização das relações. Acredita-se que *tu* passa a ocupar, paulatinamente, ao longo do período sob análise, os domínios que antes pertenciam às formas de tratamento de base nominal. Esses resultados contrariam a concepção de Cintra (1986[1972]), de que, em Portugal, predomina o uso dos tratamentos de base nominal. Entretanto, somente quando se analisar detidamente o comportamento das formas em função das relações sociais estabelecidas, considerações mais conclusivas a esse respeito poderão ser tecidas.

Já com relação à forma *você*, nota-se, de antemão, um comportamento bastante semelhante entre o PB e o PE na segunda metade do século XIX — nos dois quartos analisados e nas duas amostras, as frequências estão em sempre abaixo dos 10%. Posteriormente, no século XX, observa-se uma grande diferenciação entre essas variedades. Por um lado, na amostra brasileira, é possível encontrar uma frequência próxima a 30% relacionada à forma *você* já no primeiro quarto do século e um crescimento substancial ao longo nos quartos seguintes (54% e 80%), até atingir índices próximos a 90% no final do século, ratificando os resultados de Paredes Silva (2000). Por outro, na amostra portuguesa, as frequências de uso dessa forma se encontram sempre abaixo dos 20%.

No que diz respeito ao emprego de *vós*, nota-se que o uso desse pronome é bastante esporádico e pouco significativo nas duas variedades, visto que, segundo Cintra (1986 [1972]), já no século XVIII era considerado um traço arcaizante.

Quando se analisa comparativamente o emprego da estratégia *o(a) senhor(a)*, dois fatos mostram-se bastante relevantes. O primeiro deles diz respeito aos índices globais dessa forma serem muito mais elevados no PB do que no PE. Nesse caso, deve-se ressaltar as frequências encontradas no último quarto

do século XIX e na primeira metade do século XX, que, na amostra brasileira, estão acima do dobro das identificadas na amostra portuguesa. Uma explicação para esses resultados se encontra na diversidade de formas nominais encontradas no PE, o que acarreta uma maior quantidade de opções de escolha para os falantes portugueses, fazendo com que os índices de uso de formas nominais se diluam em três ou mais estratégias — fato que não ocorre no PB. O segundo fato relevante relacionado à análise comparativa do emprego de *o(a) senhor(a)* diz respeito à significativa queda nos índices dessa forma no PB da segunda metade do século XX. Acredita-se que essa queda esteja relacionada a mudanças no interior das relações sociais. Tal fato será investigado mais detidamente quando se analisarem os resultados diluídos entre as relações sociais estabelecidas.

Ao comparar os índices de formas nominais distintas a *o(a) senhor(a)* entre as variedades brasileira e portuguesa, destaca-se a maior frequência dessas estratégias no PE. Em primeiro lugar, na primeira metade do século XIX, pode-se verificar que a forma *V.M.* (e variantes) concorre ‘de igual para igual’ com o pronome *tu*, com índices na casa dos 40%, aparecendo, posteriormente, de modo isolado com frequências sempre abaixo dos 5%. Em segundo lugar, a forma *V.Ex.*, que sequer figura na variedade brasileira, no PE, apesar de apresentar valores quase sempre abaixo dos 13%, aparece tanto no século XIX quanto no século XX. Em terceiro lugar, há o emprego de SNs no PE ao longo de quase toda a amostra, embora suas frequências estejam sempre abaixo dos 25%, são sempre mais elevadas que o uso dessa estratégia no PB. Com relação aos SNs, é interessante comentar o desaparecimento dessas estratégias no último quarto do século XX nas duas amostras em estudo. Nesse caso, por um lado, no PB, confirma-se a tendência em curso desde o último quarto do século XIX de frequências muito baixas para o uso dessas formas, por outro, no PE, não é possível afirmar que essas estratégias tenham deixado de ser usadas, visto que apenas uma peça foi analisada neste período.

### 4.3. A distribuição geral das formas por relações sociais

#### 4.3.1. A distribuição das formas nas relações simétricas solidárias

As relações simétricas são aquelas em que nenhum dos participantes da situação comunicativa exerce poder sobre o outro, orientando-se pelo parâmetro da solidariedade, segundo Brown e Gilman (2003[1960]). São solidárias quando os participantes apresentam intimidade entre si. Nas obras que constituem as amostras, as relações simétricas solidárias foram encontradas na interação entre casais, irmãos e amigos. A seguir, encontra-se a distribuição dos 2884 dados das relações simétricas solidárias na amostra

Tabela 2. A distribuição dos dados nas relações simétricas solidárias

FT	TU		VOCÊ		VÓS		SR.		SN		TOTAL	
	PB	PE	PB	PE	PB	PE	PB	PE	PB	PE	PB	PE
VAR.												
XIX (1/2)	-	21/27 (78%)	-	-	-	6/27 (22%)	-	-	-	-	-	27
XIX (3)	218/309 (71%)	85/93 (91%)	30/309 (10%)	-	-	-	56/309 (18%)	7/93 (8%)	5/309 (2%)	1/93 (1%)	309	93
XIX (4)	1/6 (17%)	69/78 (88%)	5/6 (83%)	8/78 (10%)	-	-	-	1/78 (1%)	-	-	6	78
XX (1)	92/121 (76%)	246/250 (98%)	8/121 (7%)	4/250 (2%)	-	-	21/121 (17%)	-	-	-	121	250
XX (2)	1/241 (0%)	26/26 (100%)	227/241 (94%)	-	-	-	13/241 (5%)	-	-	-	241	26
XX (3)	5/496 (1%)	426/426 (100%)	491/496 (99%)	-	-	-	-	-	-	-	496	426
XX (4)	-	232/237 (98%)	571/575 (99%)	5/237 (2%)	-	-	4/575 (1%)	-	-	-	575	237
TOTAL	317/ 1748 (18%)	1105/ 1136 (97%)	1332/ 1748 (76%)	17/ 1136 (1%)	-	6/ 1136 (1%)	94/ 1748 (5%)	8/ 1136 (1%)	5/ 1748 (0%)	1/ 1136 (0%)	1748	1136



Inicialmente, é necessário esclarecer que não foram encontrados dados relativos a relações simétricas solidárias, na amostra portuguesa, na obra de 1925. Tal fato deve-se fundamentalmente à natureza dessa obra, que se desenvolve exclusivamente pela interação entre dois personagens, que estabelecem entre si uma relação simétrica não-solidária.

No século XIX e no primeiro quarto do século XX, observa-se uma quantidade maior de formas para o tratamento entre iguais íntimos principalmente no PB. Na amostra brasileira, por um lado, observa-se, nesse tipo de relação, a variação entre as formas pronominais, com uma forte tendência ao emprego de *tu*, embora no último quarto do século *você* predomine. O predomínio de *você*, entretanto, corresponde à frequência desse item em apenas uma peça de Machado de Assis, que apresenta, apenas 6 ocorrências nesse tipo de relação. Para entender esses dados na obra de Machado de Assis, são muito interessantes as constatações de Biderman (1972), que, ao analisar cartas pessoais também desse autor, verificou o emprego do pronome *tu* com os íntimos até o final da década de 1870, ao passo que, nas últimas décadas do século XIX, notou que se empregava quase que exclusivamente a forma *você* para os mesmos. Ressalte-se também a frequência de 17% da forma *o(a) senhor(a)*, que corresponde ao tratamento entre noivos, na peça *Quebranto*, de 1908, em que há uma grande diferença de idade da noiva em relação ao noivo. Na amostra portuguesa, por outro lado, verifica-se, um forte predomínio do pronome *tu*, embora formas como *vós*, *o(a) senhor(a)* e SNs também apareçam com frequências bem menores.

Já a partir do segundo quarto do século XX, enquanto no PB, o pronome *tu* praticamente desaparece, e há uma forte predominância de *você*, no PE, a forma *tu* aparece quase na totalidade dos casos.

Em suma, a TD evocada, no PB, em situações simétricas solidárias até o terceiro quarto do século XIX é *tu*, no último quarto do século XIX e no primeiro do século XX, há uma variação entre *tu* e *você*, e, a partir da década de 1930, *você* se consolida como a forma de tratamento a ser evocada nesses contextos. Confirmam-se, assim, os resultados obtidos por Paredes Silva (2000), Soto (2001), Salles (2001) e Lopes & Duarte (2003), que afirmam que a forma de tratamento mais produtiva nas relações simétricas solidárias no século XIX é o pronome *tu*, bem como o de Paredes Silva (2000), que aponta *você* como a estratégia mais produtiva nessas relações a partir da década de 1930.

Em contrapartida, na amostra portuguesa, a TD evocada tanto no século XIX quanto no XX é o pronome *tu*, que predomina em todos os

períodos analisados, chegando a se apresentar como (praticamente) categórico em onze das treze peças em que se identificaram relações simétricas solidárias. Tal fato evidencia que, ao contrário do que foi constatado na amostra brasileira, não ocorreu nenhuma mudança no sistema de formas de tratamento no domínio das relações simétricas solidárias na amostra portuguesa.

#### 4.3.2. A distribuição das formas nas relações simétricas não-solidárias

Entende-se por relações simétricas não-solidárias todos os tipos de relação constituído pelas interações entre desconhecidos bem como entre conhecidos, não-amigos, sem nenhum vínculo familiar. Agruparam-se também sob esse rótulo as relações entre sogro(a) e genro/nora. Tal opção se fez, uma vez que, ao contrário das demais relações de parentesco, nesta não parece, muitas vezes, nas peças, se desenvolver uma intimidade recíproca.

É fundamental sublinhar que se crê, neste estudo, que, entre todos os tipos de relação, a análise desse tipo seja o mais difícil, uma vez que as 'regras' que regem o uso de formas de tratamento entre não-íntimos são bastante 'maleáveis'.

Antes de iniciar a análise, é importante esclarecer que, na amostra portuguesa, não se identificaram relações simétricas não-solidárias nas obras de 1825, 1895, 1902, 1931 e 1995. Analisou-se um total de 2156 dados distribuídos da seguinte forma na amostra.



Com relação ao PB constata-se, durante quase todo o período sob estudo, com exceção feita ao último quarto do século XX, que a principal TD evocada nas relações simétricas não-solidárias é a forma *o(a) senhor(a)*, que predomina sobre as formas pronominais *tu*, *você* e *vós* e sobre os SNs. Já com relação ao PE, a forma *o(a) senhor(a)* prepondera em três períodos, os dois últimos quartos do século XIX e o segundo e terceiro quartos do século XX, sendo a forma *V.M.* a principal estratégia de tratamento entre iguais não-íntimos na primeira metade do século XIX.

Cabe ressaltar o emprego da forma *V.M.* e variantes ao lado da forma mais nova *você*. O emprego de *você*, no entanto, segundo Lešková (2012), na atualidade, não é utilizado por de mais de 60% dos portugueses e os que o utilizam, fazem-no para indicar respeito. É importante sublinhar os significativos usos de SNs, que praticamente não ocorrem no PB, mas se mostram muito usuais no PE. Como já foi descrito, não se observou esse tipo de relação no PE no último quarto do século XX.

É fundamental sublinhar também que, enquanto no PB há um crescente aumento da forma *você*, chegando a 77% no último quarto do século XX, no PE, o emprego de *tu*, pronome predileto entre os portugueses apresenta índices sempre abaixo dos 30%. Isso pode indicar que, no domínio simétrico não-solidário, as relações tenham se tornado mais flexíveis no PB do que no PE.

#### 4.3.3. A distribuição das formas nas relações assimétricas descendentes

As relações assimétricas são aquelas em que há poder envolvido na interação entre os interlocutores. As relações assimétricas descendentes são as que se desenvolvem na interação entre um emissor que detém poder sobre um receptor, ou seja, quem fala está em uma posição dominante em relação a seu interlocutor. Nas amostras sob análise, o tratamento assimétrico descendente é observado na fala do pai dirigida a seu filho; do tio, a seu sobrinho; do patrão, a seu empregado, entre outras.

De antemão, é necessário esclarecer que não foram identificados, na amostra brasileira, dados de relações assimétricas descendentes nas obras “Amor com amor se paga”, de 1870, e “Comunhão de bens”, de 1980. Nas obras da amostra portuguesa, não foram identificadas relações assimétricas nas obras “Viva da Costa”, de 1925, e “Os outros”, de 1965. Nas demais obras, coletou-se um total de 1310 dados distribuídos da seguinte forma:

Tabela 4. A distribuição dos dados nas relações assimétricas descendentes

FT	TU		VOCÊ		V.M. (e variantes)		SR.		SN		TOTAL	
	PB	PE	PB	PE	PB	PE	PB	PE	PB	PE	PB	PE
VAR.												
XIX (1/2)	-	46/ 52 (88%)	-	-	-	4/ 52 (8%)	-	-	-	2/ 52 (4%)	-	52 (4%)
XIX (3)	140/ 155 (90%)	128/ 154 (83%)	8/ 155 (5%)	8/ 154 (5%)	-	-	1/ 155 (1%)	2/ 154 (1%)	6/ 155 (4%)	16/ 154 (10%)	155	154
XIX (4)	-	20/ 20 (100%)	-	-	-	-	3/ 4 (75%)	-	1/ 4 (25%)	-	4	20
XX (1)	77/ 224 (34%)	71/ 71 (100%)	120/ 224 (54%)	-	-	-	27/ 224 (12%)	-	-	-	224	71
XX (2)	-	57/ 57 (100%)	46/ 49 (94%)	-	-	-	3/ 49 (6%)	-	-	-	49	57
XX (3)	-	211/ 233 (91%)	87/ 97 (90%)	19/ 233 (8%)	-	2/ 233 (1%)	10/ 97 (10%)	-	-	1/ 233 (0%)	97	233
XX (4)	-	81/ 84 (96%)	110/ 110 (100%)	3/ 84 (4%)	-	-	-	-	-	-	110	84
TOTAL	217/ 639 (34%)	614/ 671 (92%)	371/ 639 (58%)	30/ 671 (4%)	-	6/ 671 (1%)	44/ 639 (7%)	2/ 671 (0%)	7/ 639 (1%)	19/ 671 (3%)	639	671

Com base nos resultados, pode-se afirmar, inicialmente, que, no tratamento do superior ao inferior, tanto no PB quanto no PE, predomina o uso de estratégias pronominais. Na amostra brasileira, de um modo geral, é possível constatar dois comportamentos distintos. Por um lado, observa-se, inicialmente, nas obras do terceiro quarto do século XIX, o emprego predominante de *tu*; posteriormente, já no século XX, nota-se seu declínio até seu desaparecimento a partir da obra do segundo quarto desse século. Por outro lado, a forma *você*, que apresenta índices insignificantes no século XIX, a partir do século XX predomina em todos os períodos, chegando a ser categórica no último quarto desse século nesse tipo de relação.

Já na amostra portuguesa, ao analisar as formas de tratamento nas falas de personagens ‘superiores’ dirigidas a ‘inferiores’, observou-se o uso praticamente categórico de *tu* durante todo o período sob análise.

Ressalte-se que não se tecerá considerações sobre o último quarto do século XIX no PB por se tratar de apenas quatro dados em uma peça específica.

De um modo geral, o comportamento dos dados nas relações assimétricas descendentes se assemelha bastante ao comportamento nas relações simétricas solidárias.

#### 4.3.4. A distribuição das formas nas relações assimétricas ascendentes

O tratamento ascendente corresponde àquele que em que o emissor é ‘hierarquicamente’ inferior ao receptor. Tal tipo de relação é perceptível, no *corpus* analisado, na fala de filho para o pai, do sobrinho para o tio, do empregado para o patrão, do escravo para seu senhor, entre outras. É importante ressaltar que não foram identificadas relações assimétricas na peça de 1980, na amostra brasileira, e nas obras de 1925 e 1965 na amostra portuguesa. O total de 795 dados distribuíram-se do seguinte modo pelas amostras:

Tabela 5. A distribuição dos dados nas relações assimétricas ascendentes

FT	TU		VOCÊ		VÓS		V.M. (e varian- tes)		V.EX.		SR.		SN		TOTAL	
	PB	PE	PB	PE	PB	PE	PB	PE	PB	PE	PB	PE	PB	PE	PB	PE
VAR.																
XIX (1/2)	-	-	-	-	-	1/ 43 (2%)	-	37/ 43 (86%)	-	-	-	3/ 43 (7%)	-	2/ 43 (5%)	-	43
XIX (3)	1/ 128 (1%)	-	-	-	-	-	22/ 128 (17%)	3/ 38 (8%)	-	17/ 38 (45%)	34/ 128 (27%)	2/ 38 (5%)	71/ 128 (55%)	16/ 38 (42%)	128	38
XIX (4)	1/ 1 (100%)	42/ 45 (93%)	-	-	-	-	-	-	-	1/ 45 (2%)	-	-	-	2/ 45 (24%)	1	45
XX (1)	-	21/ 42 (50%)	4/ 100 (4%)	9/ 42 (21%)	3/ 100 (3%)	-	-	-	-	-	82/ 100 (82%)	-	11/ 100 (11%)	12/ 42 (29%)	100	42
XX (2)	-	14/ 28 (50%)	3/ 70 (4%)	-	-	-	-	-	-	1/ 28 (4%)	66/ 70 (94%)	-	1/ 70 (1%)	13/ 28 (46%)	70	28
XX (3)	-	44/ 98 (45%)	50/ 123 (23%)	-	-	-	-	-	-	-	73/ 123 (77%)	2/ 98 (2%)	-	52/ 98 (53%)	123	98
XX (4)	2/ 69 (2%)	10/ 10 (100%)	8/ 83 (37%)	-	-	-	-	-	-	-	59/ 83 (61%)	-	-	-	83	10
TO- TAL	4/ 491 (1%)	131/ 304 (43%)	65/ 491 (13%)	9/ 304 (3%)	3/ 491 (1%)	1/ 304 (1%)	22/ 491 (4%)	40/ 304 (13%)	-	19/ 304 (6%)	314/ 491 (64%)	7/ 304 (2%)	83/ 491 (17%)	97/ 304 (32%)	491	304

Nas obras do século XIX, na amostra brasileira, o tratamento assimétrico ascendente é encontrado predominantemente nas obras do terceiro quarto do século XIX, sendo o resultado do último quarto do século XX insignificante por se tratar de apenas um dado. O tratamento, no terceiro quarto do século XIX, é feito preponderantemente através de SNs, também sendo encontradas, em índices mais baixos, mas relativamente significativos, formas como *V. Ex* e *V.M.* e variantes.

Na amostra portuguesa do século XIX, cabe ressaltar a grande quantidade de tipos de estratégias encontradas para a referência a um interlocutor que apresenta poder frente ao falante. Uma das explicações possíveis para a diversidade de formas encontradas na amostra europeia é a significativa estratificação encontrada na sociedade portuguesa. Na primeira metade do século XIX predomina, com elevado índice, a forma *V.M.* e variantes. Já no terceiro quarto do século XIX preponderam os usos de *V.Ex.* e SNs, com índices bastante próximos aos 45%. No último quarto do século XIX, a forma predominante é *tu*, possivelmente pelos detentores de poder na relação social, no caso pais e sogros, apresentarem-se com uma visão negativa para a sociedade — o pai que negocia a filha e a mãe que é solteira.

Já nas obras do século XX, na amostra brasileira, os índices de *o(a)senhor(a)* predominam durante todo o século XX, também podendo ser encontradas, com frequência significativa, na segunda metade do século XX, a forma *você* (23% e 37%, no terceiro e quartos do século XX, respectivamente). Ressalta-se os quatro dados de *tu* em seu emprego não padrão, ou seja, relacionado a formas verbais de terceira pessoa encontrado no último quarto do século XX.

Na amostra portuguesa, nos dois primeiros e no último quarto do século XX, há um elevado emprego de *tu*. A explicação para tal fato pode estar na natureza da temática familiar dessas peças. Nesse caso, não se pode deixar de comentar que a relação entre pais e filhos, netos e avós tornou-se aparentemente mais próxima. Tal fato é observado também na Espanha, já em 1980, na descrição de Alba de Diego e Sánchez Lobato, ao investigar a fala de jovens, como uma mudança consolidada.

Los resultados expuestos confirman la evolución de los tratamientos asimétricos hacia simétricos, en los que predomina la solidaridad. Se ha pasado así a una forma más democrática e igualitaria en las relaciones familiares. [...] En esta línea creciente del *tuteo* hemos de interpretar la extensión que ha tomado *tu* en fórmulas de tratamiento que hasta hace poco tiempo, y nuestra conciencia lingüística así lo atestigua, eran campo reservado al *Ud.* (Alba de Diego & Sánchez 1980, p.113)



É fundamental também comentar que o uso de SNs, que são, por sua natureza, heterogêneas, a partir da segunda metade do século XIX, possui índices superiores a 24%, com exceção feita ao último período em análise, em que o uso de *tu* é categórico.

## 5. Considerações finais

Com base nas discussões desenvolvidas ao longo deste estudo e na análise dos resultados obtidos na distribuição das formas de tratamento ao interlocutor, nas vinte e nove peças teatrais, foi possível traçar um interessante panorama do sistema de tratamento ao interlocutor no PB e no PE dos séculos XIX e XX.

É fundamental reiterar que, neste estudo, se consideram as formas de tratamento ao interlocutor *Tradições Discursivas*, visto que se entende que estas são formas linguísticas evocadas em situações comunicativas específicas. Acredita-se também que mudanças tanto na semântica das formas quanto na natureza das relações sociais tenham gerado contextos em que duas ou mais estratégias competem em um mesmo domínio. Essa competição é o que tradicionalmente se conhece como *variação linguística*.

Na análise dos dados., constatou-se que, ao longo dos séculos XIX e XX, em ambas as variedades, há uma tendência à simplificação dos sistemas de tratamento ao interlocutor.

Comparando as duas variedades, em linhas gerais, é possível afirmar que há dois momentos distintos. O primeiro deles, no século XIX e primeiro quarto do século XX, em que se verifica uma certa similaridade entre os sistemas de tratamento do PB e do PE, em que o emprego do pronome *tu* se mostra quase sempre predominante e há, com frequências mais baixas, outras formas — *você, vós, V.M., V. Ex.* (exclusivamente no PE), *o(a) senhor(a)* e SNs.

O segundo corresponde ao período compreendido a partir do segundo quarto do século XX. Nesse período, existe uma progressiva diferenciação entre as duas variedades da língua. Na amostra brasileira, o sistema de tratamento parece se restringir às formas *você* e *o(a) senhor(a)*, com o crescente predomínio daquela sobre esta. Já na amostra portuguesa, essa forma pronominal não só é a predominante em todos os períodos sob análise, como paulatinamente aumentam seus índices de uso, embora ainda haja uma variedade de formas muito mais ampla no PE do que no PB.

Com relação à distribuição pelas relações sociais estabelecidas, nota-se uma similaridade entre os resultados das relações simétricas solidárias e

das relações assimétricas descendentes. Ao contrário do esperado, as relações simétricas não-solidárias e as relações assimétricas descendentes apresentaram comportamentos distintos.

Com relação ao PB, por um lado, observou-se que o sistema de tratamento ao interlocutor, nas relações simétricas solidárias e assimétricas descendentes, vivencia três momentos distintos:

- a) No primeiro momento, no século XIX e no primeiro quarto do século XX, o tratamento se dá predominantemente através do pronome *tu*, embora este possa ter dado lugar a *você* já na passagem do século XIX para XX, como se verifica na obra de Machado de Assis, de 1896, e nos apontamentos dos trabalhos de Biderman (1972) e Paredes Silva (2000).
- b) No segundo momento, entre as décadas de entre o segundo quarto do século XIX e o último quarto do século XX, a referência ao interlocutor ocorre praticamente de maneira categórica com a forma *você*.
- c) No final do século XX, é possível verificar, de maneira tímida, a variação entre as formas *você* e *tu*, esta em seu emprego *não-padrão*, ou seja, acompanhada pela flexão do verbo na 3ª pessoa.

Por outro lado, nas relações simétricas não-solidárias, observa-se a preponderância da forma *o(a) senhor(a)* em praticamente todos os períodos sob análise, exceção feita ao último quarto do século XX, em que a forma de maior frequência é *você*. A escalada das frequências dessa forma pronominal, aliás, é um dos fatos que merece destaque. Se, no primeiro quarto do século XX, correspondia a cerca de 20% do total de ocorrências encontradas nas relações simétricas não-solidárias, a partir do último quarto do século XIX, atinge um percentual próximo a 80%. Tal fato pode indicar que as relações marcadas pelo distanciamento entre os interlocutores caminham em direção à intimidade. Resultados semelhantes são identificados na análise das relações assimétricas ascendentes. Assim como nas relações simétricas não-solidárias, a forma *o(a) senhor(a)* é mais produtiva em quase todos os períodos sob investigação.

Com relação ao PE, de um modo geral, observa-se que o emprego de *tu* é praticamente categórico nas relações simétricas solidárias e nas relações assimétricas descendentes ao longo de todo o período sob análise. Já nas relações simétricas não-solidárias, verifica-se uma multiplicidade de formas que só é possível explicar com uma análise pormenorizada das situações interacionais. Por fim, nas relações assimétricas ascendentes, também se constata uma grande diversidade de formas, mas, ao contrário do que se observa nas relações simétricas não-solidárias, nota-se um aumento, por

vezes, progressivo, nos índices de *tu*, o que indica que as relações de poder podem estar se flexibilizando.

De modo geral, verifica-se que, ao longo dos dois últimos séculos, as alterações ocorridas no quadro do tratamento do PB se mostram muito mais intensas do que as observadas no PE. As mudanças identificadas na variedade do PE estão ligadas, de modo quase que exclusivo, a modificações nas relações sociais, que também se constata na variedade brasileira. Essas mudanças estão relacionadas, sobretudo, às transformações vivenciadas no interior das sociedades que, a partir, principalmente, de meados do século XX, tendem a flexibilizar as relações de poder. Tal fato fica bastante evidente com o aumento nos índices de uso de estratégias pronominais, nas duas variedades, que se sobrepõem ao emprego de formas de base nominal. Estas aparentemente sobrevivem como formas cristalizadas pela Tradição, e não por sua semântica de distanciamento.

## Referências

- Alba de Diego, V. & Sánchez Lobato, J. (1980). Tratamiento y juventud en la lengua hablada. *Boletín de la Real Academia Española* LX/ CCXIX, 45–130.
- Biderman, M. T. C. (1972). Formas de Tratamento e Estruturas Sociais. *Alfa*, 18/19, 339–381. São Paulo: FFCL de Marília.
- Brown, R.; Gilman, A. (2003 [1960]). The pronouns of power and solidarity. In C. B. Paulston & G. R. Tucker (Eds.), *Sociolinguistics: The essential readings* (pp. 156–176). Blackwell.
- Cintra, L. F. L. (1986). *Sobre “formas de tratamento” na língua portuguesa*. (2ª ed.) Lisboa: Livros Horizonte.
- Faraco, C. A. (1996). O tratamento *você* em português: uma abordagem histórica. *Fragmenta 13, Publicação do Curso de Pós-Graduação em Letras da UFPR*. Curitiba: Editora UFPR.
- Goffman, E. (1980) A elaboração da face. Uma análise dos elementos rituais na interação verbal. In S. Figueira (Ed.), *Psicanálise e ciências sociais* (pp. 76–114). Rio de Janeiro: Livraria Francisco Alves.
- Goffman, E. (2007). *As representações do eu na vida cotidiana*. (14ª ed.) Petrópolis: Vozes.
- Kabatek, J. (2006). Tradições discursivas e mudança linguística. In T. Lobo, I. Ribeiro, Z. Carneiro & N. Almeida (Eds.), *Para a história do português brasileiro. VI: Novos dados, novas análises* (Tomo II, pp. 505–527). Salvador, Brasil: EDUFBA.
- Kerbrat-Orecchioni, C. (2006). *Análise da conversação. Princípios e Métodos*. São Paulo: Parábola Editorial.

- Lopes, C. R. dos S., Couto, L. R. & Duarte, M. E. L. (2005) Como as pessoas se tratam no cinema latino-americano: Análise de formas de tratamento em roteiros de três países. *Memórias – XIV Congresso Internacional da ALFAL*, vol. 1.
- Lopes, C. R. dos S. & Duarte, M. E. L. (2003). De ‘Vossa Mercê’ a ‘você’: A análise pronominalização de nominais em peças brasileiras e portuguesas setecentistas e oitocentistas. In S. F. Brandão & M. A. Mota (Eds.), *Análise contrastiva de variedades do português. Primeiros estudos* (vol. 1, pp. 61–76). Rio de Janeiro: In-Fólio.
- Lopes, C. R. dos S. & Machado, A. C. M. (2005). Tradição e inovação: indícios do sincretismo entre segunda e terceira pessoas nas cartas dos avós. In C. R. dos S. Lopes (Ed.), *Norma brasileira em construção: Fatos lingüísticos em cartas pessoais do século XIX* (Pós-Graduação, FAPERJ).
- Loregian-Penkal, L. (2004). *(Re)análise da referência de segunda pessoa na fala da região sul* (Tese de doutoramento, Universidade Federal do Paraná).
- Lešková, J. (2012). *As formas de tratamento em Português Europeu* (Tese de doutoramento, Univerzita Palackého v Olomouci). Disponível em: <[https://theses.cz/id/lfal0x/diplomov\\_prce.pdf](https://theses.cz/id/lfal0x/diplomov_prce.pdf)>
- Machado, A. C. M. (2006). *A implementação de você no quadro pronominal: as estratégias de referência ao interlocutor em peças teatrais no século XX* (Diss. de mestrado, UFRJ).
- Menon, O. P. da S.; Loregian-Penkal, L. (2002). Variação no indivíduo e na comunidade: tu/você no sul do Brasil. In P. Vandresen (Ed.), *Variação e mudança no Português falado da região sul*. Pelotas: Educat.
- Paredes Silva, V. L. (1999). *O percurso da variação na referência à segunda pessoa no português carioca*. Relatório final de pesquisa apresentado ao CNPq (p. 35). Rio de Janeiro: UFRJ, Mimeo.
- Paredes Silva, V.L. (2000). A distribuição dos pronomes de segunda pessoa do singular na fala carioca ao longo do século XX. *II Congresso Nacional da Abralin (CD-rom)*.
- Paredes Silva, V. L. (2003). O retorno do pronome *tu* à fala carioca. In C. Roncarati, J. Abraçado (Eds.), *Português brasileiro – contato lingüístico, heterogeneidade e história* (pp. 160–169). (1ª ed.) Rio de Janeiro: 7letras/FAPERJ.
- Robinson, W. P. (1977 [1972]). *Linguagem e comportamento social*. J. Martins (Trad.). São Paulo: Cultrix.
- Rumeu, M. C. de B. (2004). *Para uma história do português no Brasil: formas pronominais e nominais de tratamento em cartas setecentistas e oitocentistas* (Diss. de mestrado, UFRJ).
- Rumeu, M. C. de B. (2008). *A implementação do ‘você’ no português brasileiro oitocentista e novecentista: um estudo de painel* (Tese de doutoramento, UFRJ).
- Sales, M. (2001). *Pronomes de tratamento do interlocutor no português brasileiro: um estudo de pragmática histórica* (Tese de doutoramento, USP).

- Soto, E. (2001). *Variação/Mudança do pronome de tratamento alocutivo: uma análise enunciativa em cartas brasileiras* (Tese de doutoramento, Universidade Estadual Paulista 'Julio de Mesquita Filho').
- Weinreich, U., Labov, W., Herzog, M. (2006). *Fundamentos empíricos para uma teoria da mudança lingüística*. M. Bagno (Trad.). Revisão Técnica de Carlos Alberto Faraco. Posfácio de Maria da Conceição e Maria Eugênia Lamoglia Duarte. São Paulo: Parábola Editorial.

### As obras que compõem a amostra

- Alencar, J. [1857]. *O demônio familiar*. Disponível em: <www.dominiopublico.com.br>.
- Alves, V. (2006 [1925]). Viva Da Costa. In L. Rebello (Ed.), *O teatro português em um acto (século XX)*. Lisboa: Imprensa Nacional – Casa da Moeda.
- Amaral, M. A. de. (1995). Intensa Magia. *SBAT*, 493/494, 45. Mimeo.
- Araújo, A. (1980). Comunhão de bens. *SBAT*, 436, 78. Mimeo.
- Barbosa, M. C. & Falabella, M. (2004 [2003]). Síndromes. In M. C. Barbosa & Falabella, M. (Eds.), *Querido Mundo e outras peças*. Rio de Janeiro: Lacerda.
- Bloch, P. (1973 [1952]). *Dona Xepa*. Rio de Janeiro: Serviço Nacional de Teatro (Coleção Dramaturgia Brasileira).
- Bordalo, J. (2003 [1825]). O Beato Ardiloso. In L. Rebello (Ed.), *O teatro português em um acto (1800–1899)*. Lisboa, Portugal: Imprensa Nacional – Casa da Moeda.
- Chaves, P. C. (1866). *O Ferro Velho*. Lisboa: Typographia da Viuva Pires Marinho.
- Chagas, M. (2003) Quem desdenha.... In L. Rebello (Ed.), *O teatro português em um acto: volume I 1800–1899*. Lisboa: Imprensa Nacional – Casa da Moeda. Escrita em 1874.
- Coelho Netto, H. M. (1957 [1908]). Quebranto. In *Revista de Teatro da SBAT*, 295. Rio de Janeiro.
- Curto, R. (2006 [1931]). Três gerações. In L. Rebello (Ed.), *O teatro português em um acto (século XX)*. Lisboa: Imprensa Nacional – Casa da Moeda.
- Dinis, J. (2003 [1858]). Similia Similibus. In L. Rebello (Ed.), *O teatro português em um acto (1800–1899)*. Lisboa: Imprensa Nacional – Casa da Moeda.
- Figueiredo, T. (2003 [1958]). *O homem do quiosque*. In T. Figueiredo (Ed.), *Teatro*. Lisboa: Imprensa Nacional – Casa da Moeda.
- Fortuna, R. (2003 [1819]). As astúcias de Zanzingarra. In Rebello, L (Ed.), *O teatro português em um acto (1800–1899)*. Lisboa: Imprensa Nacional – Casa da Moeda.
- França Junior, J.J. [1870] *Amor com amor se paga*. Disponível em: <www.dominiopublico.com.br>.
- França Junior, J.J. [1870]. *O defeito de família*. Disponível em: <www.dominiopublico.com.br>.

- Gill, G. (1964 [1962]). *Toda donzela tem um pai que é uma fera*. Rio de Janeiro: Tempo Brasileiro.
- Gonzaga, A. (1937). *O hóspede do quarto número 2*. Rio de Janeiro, Brasil: SBAT - Mímeo.
- Machado De Assis, J. (1896). *Não Consultes Médico*. Disponível em: <[www.dominiopublico.com.br](http://www.dominiopublico.com.br)>.
- Melo, J. (1995) António, um rapaz de Lisboa. Lisboa: Edições Cotovia.
- Mesquita, M. (2003 [1895]). Fim de penitência. In L. Rebello (Ed.), *O teatro português em um acto (1800–1899)*. Lisboa: Imprensa Nacional – Casa da Moeda.
- Mesquita, M. (2006 [1902]). O tio Pedro. In L. Rebello (Ed.), *O teatro português em um acto (século XX)*. Lisboa: Imprensa Nacional – Casa da Moeda.
- Ogando, A. (2006 [1934]). A prima tança. *O teatro português em um acto (século XX)*. Lisboa: Imprensa Nacional da Casa da Moeda.
- Pena, M. [1846] *Os Ciúmes de um Pedestre*. Disponível em: <[www.dominiopublico.com.br](http://www.dominiopublico.com.br)>
- Rebello, L. (1999 [1956]). É urgente o amor. In L. Rebello (Ed.), *Todo o teatro*. Lisboa: Imprensa Nacional – Casa da Moeda.
- Rocha, A. (1979 [1972]). O genro que era nora. *Revista de Teatro da SBAT*. Rio de Janeiro.
- Tojeiro, G. (1966 [1918]). O Simpático Jeremias. *Revista de Teatro da SBAT*. Rio de Janeiro.
- Sampaio, J. (1974 [1965]). Os outros. In: *Seis Peças*. Lisboa: Plátano Editora.

[recebido em 31 de janeiro de 2018 e aceite para publicação em 14 de janeiro de 2019]



# THE ROLE OF PRAGMATIC MARKERS IN ACADEMIC SPOKEN INTERLANGUAGE: A CORPUS-BASED STUDY OF A GROUP OF BRAZILIAN EFL UNIVERSITY STUDENTS

O PAPEL DOS MARCADORES PRAGMÁTICOS NA  
INTERLÍNGUA FALADA ACADÊMICA:  
UM ESTUDO BASEADO EM CORPUS DE UM GRUPO DE  
ESTUDANTES UNIVERSITÁRIOS BRASILEIROS DE ILE

Bárbara Malveira Orfanò\*  
barbara.orfano@gmail.com

Ana Larissa Adorno Marciotto Oliveira\*  
adornomarciotto@gmail.com

Spencer Barbosa da Silva\*\*  
spencer.silva@ufop.edu.br

The present work addresses a group of university students of EFL (English as a Foreign Language) on how they use pragmatic markers in their oral productions. The initial hypothesis was that there would be differences both in usage and form in comparison to native speakers. In order to verify our claim, we set off to investigate two corpora: a learner oral corpus being compiled at the Federal University of Minas Gerais/Brazil and a sub-corpus from the British Academic Spoken English (BASE). While Brazilian students overuse items such as *maybe* and *just*, the data recorded in the UK displayed a more varied range of markers and multiword forms. Overall, the findings reinforce the importance of analyzing empirical data for a broader understanding of how native speakers and learners can differ in their oral academic production. The paper also sheds light on language teaching and learning in the academic setting from a pragmatic viewpoint.

**Keywords:** Learner *corpus*. Academic discourse. Politeness.

O presente trabalho aborda como um grupo de estudantes universitários de ILE (Inglês Língua Estrangeira) usam marcadores pragmáticos em suas produções

---

\* Faculdade de Letras - Universidade Federal de Minas Gerais, Brasil.

\*\* Departamento de Estatística - Universidade Federal de Ouro Preto, Brasil.



orais. O estudo consiste de dois corpora: um corpus de aprendiz sendo compilado na Universidade Federal de Minas Gerais-Brasil e um subcorpus do British Academic Spoken English (BASE). Enquanto aprendizes brasileiros sobreusam itens específicos como *just e maybe*, falantes nativos ou fluentes utilizam uma variedade maior de unidades multipalavra. Os resultados reforçam a importância da análise de dados empíricos no estudo da produção oral de aprendizes. Eles também lançam luz para o ensino e para a aprendizagem de inglês em contexto acadêmico de um ponto de vista pragmático.

**Palavras-chave:** *Corpus* de aprendiz. Discurso acadêmico. Polidez.



## 1. Introduction

Pragmatic markers, or discourse markers, allow for writers/speakers to communicate their stance or attitudes toward the information conveyed. They make room for negotiating the certainty of statements, functioning to ‘linguistically situate the intention of the writer, while priming the reader/listener to align with this intention’ (Ran 2003, p. 8).

Markers like these are commonly used to evaluate the certainty of a proposition, while concealing the author’s voice as the source of assessment and thus presenting the assertions as objective and impersonal (Biber 2006; Ran 2003). In the academic domain, pragmatic markers may also function as a strategy of negative politeness. They may contribute to soften the imposition of the research information (such as the hypotheses, the theoretical contributions, and the results) on the reader/listener, while they treat it as neutral or objective. In this sense, pragmatic markers operate as avoidance politeness strategies (Goffman 1967), in which the speaker/author prevents himself from invading the interlocutor/reader’s territory.

Politeness, taken in the comprehensive sense of speech oriented to an interactor’s public persona or ‘face’, is ubiquitous in language use (Oliveira, Cunha & Miranda 2017). It therefore meets the aim of the study we propose here, which focuses on the academic language domain. The reason for such a claim is associated with the fact that image projection is a hallmark of the academic domain. Likewise, the expression of stance in this field is also considered crucial. In this paper, we claim that the scope of pragmatic markers employed in the academic domain differ from native speakers and learners

in manifold ways. In order to verify this claim, we set off to investigate two corpora: a learner oral corpus being compiled at the Federal University of Minas Gerais/Brazil and a sub-corpus from the British Academic Spoken English (BASE).<sup>1</sup>

In the past few years, pragmatic markers or metalinguistic monitors have been under scrutiny by different researchers. Erman (2001), Aijmer (2002; 2004), McCarthy and Carter (2006) and Fung and Carter (2007) have examined pragmatic markers in written and spoken discourse. However, studies concentrating on how Brazilian university students of English use such markers in spoken interlanguage are virtually non-existent. Considering that developing students pragmatic awareness is an essential part of their academic literacy, this paper aims to shed light upon how a group of Brazilian university students use pragmatic markers in their oral presentations. Upon the implications of this use, we will also focus on the way learners interact in English with their scientific community.

The learners in this study were undergraduate students taking the course English for Academic Purposes taught at a Federal University in Brazil. In order to better understand how Brazilian university students taking this course use pragmatic markers in their spoken language, we compared the results of our learner corpus with a native speaker corpus, focusing on underuse and overuse of the most significant patterns drawn from the data. In order to achieve this, the following research questions were addressed: What are the most commonly found patterns of pragmatic markers in the two corpora analysed? In case they are different, what is the possible impact of this discrepancy, considering the demands of the academic domain?

## 2. Pragmatic Markers and the interpersonal domain

The choice for the term 'pragmatic marker' is not a fortuitous one. In fact, it follows studies stating that it is impossible to ignore text type (written or spoken) context and the relationship between interlocutors. Arguing along the same lines, this paper follows the assumption that the term 'pragmatic

---

1 The transcriptions used in this study come from the British Academic Spoken English (BASE) corpus project. The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council.

marker' suggests a high degree of context sensitivity, as is also acknowledged by Andersen (2001).

An important issue concerning pragmatic markers is their multifunctionality. Their function and use vary depending on different issues ranging from discourse markers linking units of discourse, and then being responsible for coherence, to modal items within a more interpersonal dimension. This way, as pointed out by Fung and Carter (2007, p. 414), such markers are pragmatically significant and socially sensitive.

The literature in the field shows that pragmatic markers are not an easy term to define and definitions are usually associated with different approaches to their study, as well as with the functions related to them.

Fraser (1999) defined pragmatic markers as a pragmatic class, or as lexical expressions drawn from the syntactic classes of conjunctions, adverbials, and prepositional phrases which "signal a relationship between the segment they introduce, S2, and the prior segment, S1" (Fraser 1999, p. 63). In the example below, the pragmatic marker *In spite of*, relates the explicit interpretation of S2 to a non-explicit interpretation of S1. In S2, there is an implied proposition associated with S1, which is referenced by the use of *In spite of that*.

- (1) S1 We left late.  
     S2 In spite of that, we arrived on time.  
     (Fraser 1999, p. 64)

As we can see in (1), example taken from Fraser (1999), speakers can choose between hedges and approximators when they wish to minimize (or maximize) the effect of the message being communicated. Examples such as *I think*, *maybe* and *kind of* represent some of the most common face-saving markers (Goffman 1967) used by speakers in different contexts. Erman (2001) explains that, in the case of German studies on modality, interlocutors tend to concentrate more on the expressive attitude of the speaker towards the propositional contents of the utterance. This notion is closely related to Kriwonossow's (1977, p. 187) subjective modality and to Bublitz' (1978, p. 8) emotive modality. Both perspectives are oriented towards the speaker's attitude and also to the relationship between speaker and hearer.

In a similar vein, Aijmer (2004) acknowledges the importance of pragmatic markers in the study of learners' interlanguage. According to the author, these features need to be analyzed from the students' perspective, which means that looking at how learners use pragmatic markers in their

discourse might reveal important characteristics of their oral production. Aijmer (2013) also defines the role of pragmatic markers within a general pragmatic theory that concentrates on the language user and on the relationship between meaning and context. This is the position that is followed in this study, since our main goal is to determine how learners use such markers and the implications of this use in their academic discourse.

Along the same lines, Erman (2001) states that there are two well-established functions of pragmatic markers: they can be used as monitors of discourse and as interactional features. The author observes that, as a primary function, the markers fulfill the role of textual monitors, being responsible for turning fragmented pieces of discourse into a coherent text; however the secondary function markers operate as social monitors and their main role is to promote the negotiation of meaning and discourse management, ensuring that there is an open channel between interlocutors.

Advancing in his research, Erman (2001) proposes a third function labeled 'metalinguistic monitors' or, as he prefers to name it 'metalinguistic domain'. According to the author, markers within a metalinguistic domain are usually modal and speaker-oriented, having two main roles: to emphasize the speaker's authority as to the illocutionary force of an utterance and/or to serve as a *face-saving device* (Goffman 1967). This latter role, in particular, is the view adopted in this paper. In the next section, we will briefly address the foundation of this work, concerning Politeness Theory and the notion of face-work.

### 3. Politeness Strategies, face work and image projection

The notion of face-work, as was presented by Goffman (1967) refers to "the positive social value that a person effectively claims for himself by the line that others assume he has taken during a particular contact" (Goffman 1967, p. 223). More specifically, the term face-work refers to "the actions taken by a person to make whatever he is doing consistent with face. Face-work serves to counteract "incidents" – that is, events whose effective symbolic implications threaten face" (Goffman 1967, p. 12).

Also in the realm of image projection and face-work, Brown and Levinson (1987) have shown that certain lexical, grammatical, and prosodic phenomena can only be fully explained from the perspective of sociological factors (such as power relations) and pragmatic elements (such as the principle of politeness). From this viewpoint, elements of microlinguistic nature

(in this paper adverbs, prepositional phrases and conjunctions) may act as strategies of politeness, used to mitigate, or to intensify the virtual threats inherent to Face-Threatening Acts (FTA), such as criticisms, promises, compliments, among others (see Kerbrat-Orecchioni 2006; Cunha 2015; Oliveira, Cunha and Miranda, 2017).

In Brown and Levinson's Politeness Theory (1987, p. 61), the notions of face-work and territory, taken from Goffman (1976), are revisited in order to hold the concepts of 'positive face' and 'negative face', as in (a) and (b) below:

- a) negative face: the basic claim to territories, personal preserves, rights to non-distraction, to freedom of action and to freedom from imposition;
- b) positive face: the positive self-image or 'personality' claimed by interactants.

As Orfanò (2010) also postulates, in systematizing Goffman's approach to language studies, Brown and Levinson (1987) re-formulate the concept of facework. On the one hand, it becomes more restricted, since it only corresponds to the use of linguistic procedures (and not any procedures in general) that mitigate/intensify the threat of speech acts. On the other hand, the notion is deepened, in that it encompasses the strategies used to mitigate/intensify attacks on the negative face, and no longer only attacks on the positive face.

It is on the basis of this notion of face-work that, more recently, Brown (2015, p. 326) conceptualizes politeness in these terms: "Politeness is essentially a matter of taking into account the feelings of others as to how they should be interactionally treated, including behaving in a manner that demonstrates appropriate concern for interactors". Politeness strategies are, therefore, inherent to language use in general, which embraces the academic oral domain. Taking this claim into account, in the next section, we will present the methods of data collection and analysis of this study in an attempt to verify the potential similarities and differences between the two corpora in focus, in terms of the pragmatic markers students most frequently used.

#### **4. Methodology of data collection**

This study comprises two corpora. The main corpus, the Brazilian Academic Spoken English Corpus (BRASE), consists of 20-minute oral presentations given by students taking the course English for Academic Purposes at a

Federal University in Brazil. The undergraduate students are from different degree programs and their level of proficiency ranges from B1 to C1, following the Common European Framework of Reference for Languages (CERF). At the moment of data collection, this corpus had approximately 50,000 words. The reference corpus is a sub-corpus from the British Academic Spoken English (BASE), from the Humanities area, compiled for this specific study. BASE is a corpus designed by researchers from the Centre of Applied Linguistics of the University of Warwick-UK. In total the corpus has 1,644,942 words, encompassing the areas of Arts and Humanities, Life and Medical Sciences, Physical Sciences and Social Sciences.

After transcribing, organizing and including metadata, the main corpus was submitted for analysis using the software *Wordsmith Tools 5.0*. First, a frequency list was generated and items with the potential to function as metalinguistic monitors were isolated for analysis. In order to check if the items were in fact functioning as metalinguistic monitors, concordance lines were generated and items that were fulfilling different functions, such as circumstantial adjuncts or modifiers in noun phrases, were eliminated from the analysis and a list of metalinguistic monitors were designed.

Considering that metalinguistic monitors can also be composed of more than one word, cluster lists of 2, 3 and 4 words were generated and items fulfilling a metalinguistic monitor function were selected for analysis. However, the most fruitful list was the one containing 2 word-clusters, and for this reason only this list was included in the analysis. The same procedure was carried out for the reference corpus and lists of single and 2 word clusters were generated for both corpora to be compared.

Once the lists with the metalinguistic monitors were generated, the data was submitted to a set of statistical tests (see statistical analysis section). This procedure enables the researcher to be more accurate when analysing the data, thereby avoiding misinterpretation of the results from the corpora. As an example, the Log Likelihood<sup>2</sup> test allows the researcher to determine if the items under analysis when compared to another corpus present a significant difference. It is possible to check underuse and overuse of items in relation to the reference corpus and, in this way, obtain a better account of the main characteristics of the data being analysed.

---

2 <http://ucrel.lancs.ac.uk/llwizard.html>

5. Data analysis

The investigation begins with a single item frequency search isolating the items with the potential to function as metalinguistic monitors, following O’Keeffe, McCarthy and Carter’s (2007) framework. Searches for 2 word clusters were also carried out and again any forms likely to function as metalinguistic monitors were highlighted.

From these searches the highest number of potential items resulted from the 2 word cluster search, for example, *I mean*, *I think* and *I guess*, and for that reason the present paper concentrates on these examples. The following tables demonstrate the frequency of single words.

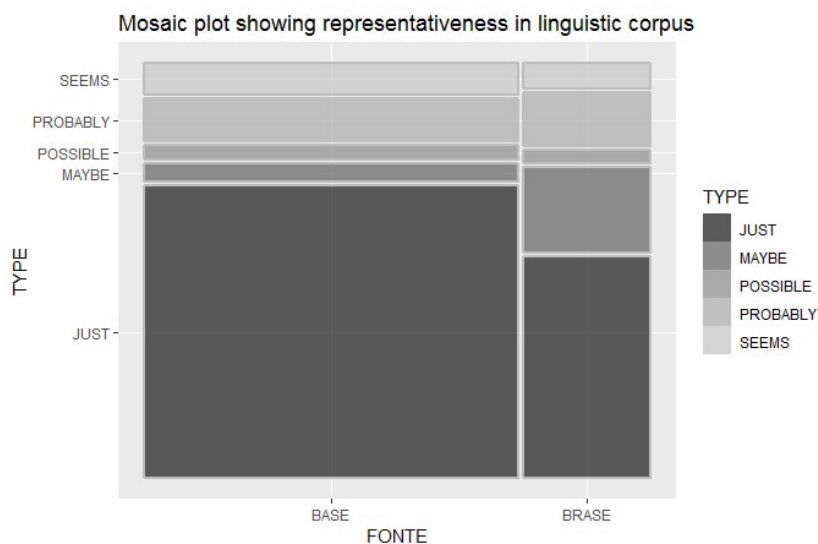
Table 1. Single words functioning as metalinguistic monitors in both corpora

BASE 150.000 words			BRASE 50.000 words		
Items	Raw freq.	Normalised Per 100.000	Item	Raw freq.	Normalised Per 100.000
JUST	401	267	JUST	103	206
ACTUALLY	316	210	MAYBE	40	80
RATHER	68	312	PROBABLY	25	50
PROBABLY	59	38	SEEMS	12	24
SUGGEST(s)	50	33	POSSIBLE	6	12
SEEM	26	46			
SEEMS	43				
MAYBE	24	16			
POSSIBLE	20	13			
POSSIBLY	13	9			
LIKELY	12	8			
SUPPOSED	11	7			
POSSIBILITY	10	6			
THINKS	10	6			
APPARENTLY	7	5			
SUPPOSEDLY	7	5			
TENDS	7	5			
INDICATIVE	6	4			
TOTAL	1.090	990	TOTAL	186	372

The first results from Table 1 indicate that there is variation in the forms used in each corpus and also in the number of forms. The frequency is higher in BASE than it is in the Brazilian group under investigation. In addition, in the British Corpus the use of items are evenly distributed, whereas in BRASE more than half of the uses are concentrated on two specific items. In order to analyze the features in more detail, we submitted the data to specific statistical tests.

### 5.1. Statistical Analysis for one-word elements

The contingency table that represents the qualitative observations extracted from the samples of the two corpora (BASE and BRASE) can be represented through the mosaic chart (Graph 1), in which each horizontally subdivided rectangle shows the proportionality of the results found in each corpus. Analysing Graph 1, we can observe a greater frequency of the words *just* and *probably* in BASE when compared to BRASE.



Graph 1. Mosaic display for sample of single words in the BASE and BRASE corpora



Table 2 below shows the results of the statistical tests carried out for the single word list. They were obtained from the comparison between the word frequency found in BRASE and BASE, via the application of Relative Risk (RR), Odds Ratio (OR) and P-value, when performing the chi-squared test in order to make proportions uniform. Besides, Log-likelihood test and P-value were also used to equalize proportions and the Confidence Interval at the 95% level was applied for the difference between proportions. These procedures were necessary due to the fact that the two corpora, BASE and BRASE, are unequal in size, each containing 150,000 words and 50,000 words respectively.

Concerning the results for Relative Risk (RR), generated by the ratio between the risk of occurrence of the word in BRASE and the risk of occurrence of that word in BASE, values higher than 1 were obtained in some situations, which means that the risk of *just* and *maybe* occurring in BRASE is higher than in BASE.

As for the Odds Ratio (OR), generated by the ratio between the odds (or “chance”) of occurrence of a word in BRASE and odds (or “chance”) of occurrence of that word in BASE, the word probably has a greater chance of being used in BRASE rather than in the BASE corpus.

The measures of the p-value generated from the chi-square test led to the finding that at the 5% significance level, a significant difference between the proportions relative to the words *just* and *maybe* can be found. In other words, these proportions differ statistically. Conversely, regarding the words *probably*, *seems*, and *possible*, the findings show that at the 5% level of significance the proportions do not differ statistically.

These same results are corroborated by the values obtained in the P-value through the test of equality of proportions, at the level of 5% of significance and by the Confidence Interval, at the level of 95% constructed for the difference in proportions (intervals containing zero). This proves that the proportions of these three words do not differ statistically. As for the values obtained from the Loglikelihood test, it is possible to conclude that the words *just* and *maybe* do not have the same proportion in BASE and in BRASE, suggesting their overuse in the Brazilian corpus. Table 2 presents the results from the tests described in this section.

**Table 2. Statistical Results of occurrence and association of simple words**

ITENS	RR	OR	P-VALUE $\chi^2$	LL	P-VALUE = PROP	IC (0,95) for D
JUST	0,817	0,771	0,01811	5,85	0,02048	[-0,081; -0,010]
MAYBE	2,5	5	0	40,03	0	[0,255; 0.495]
PROBABLY	1,191	1,271	0,3136	0,98	0,38	[-0,051; 0,146]
SEEMS	0,873	0,837	0,586	-0,31	0,697	[-0,142; 0,078]
POSSIBLE	0,923	0,9	0,8209	-0,05	1	[- 0,843; 0,146]

In order to investigate the implications of these findings in the academic discourse of learners, the next sections will deal with these metalinguistic monitors in more detail.

## 5.2. The item *just* as a sample of overuse in BRASE

Following Aijmer's (2002, p. 158) claim that the pragmatic marker *just* has procedural meaning in that it functions as a signal to the hearer to interpret the speaker's utterance as an expression of an attitude. According to McCarthy and Carter (2006), the use of *just* in oral discourse has a number of functions: It can be used for emphasis, as a particularizer, temporal meaning, limiter and as a softener or downtoner (*idem*, p. 98). Significantly important for this study are Aijmer's (2002) observations, which assign *just* to a hedging function in the realms of both positive and negative politeness. From the perspective of negative politeness, *just* functions as a downtoning hedge, modifying the face threat carried by speech acts such as assertions, suggestions, criticisms or requests (*idem*, p. 169). The analysis demonstrates that *just* is overused in BRASE indicating that learners are familiar with the hedging role fulfilled by this marker in oral discourse. However, at this point, we can claim that overusing *just* as an epistemic<sup>3</sup> marker in their oral presentations, learners are not complying with the characteristics of oral academic discourse, since they show a preference for a marker that belongs to the informal domain.

### (2) BRASE IFA 2- B1+<sup>4</sup>

3 Epistemic modality refers to the degree of commitment one has in relation to what one says.

4 IFA stands for *Inglês para Fins Acadêmicos – English for Academic Purposes*. The example was taken from students taking this course. B1 corresponds to their proficiency (intermediate) following the Common European Framework of Reference for Languages.

*the volume is basically the volume of the cylinder and the cost is **just** the cost of all materials involved... here is the result ah all the solutions here and what was in the in this two genetic algorithm ...*

Examples like the one in (2) are common in BRASE. Even though this group of students are on an intermediate level, they seem to be unaware of other markers functioning as hedges. This overreliance on one marker tends to impoverish learners’ oral academic performance. Advancing in our analysis, we decided that it would be important to look at multiword combinations considering that they are important in creating meaning and also responsible for fluency in oral discourse.

According to O’Keeffe *et al.* (2007, p. 60), “what corpora reveal is that much of our linguistic output consists of repeated multi-word units rather than just single words”. The literature has dedicated considerable attention to this issue and the terminology varies depending on the researcher and the theoretical perspective adopted in each study. Biber (1999) calls the combination of repeated words lexical bundles, O’Keeffe *et al.* (2007) refer to them as cluster units and Orfanò (2010) names them as frequent items. Due to their importance they are also analysed in this paper. Table 3 accounts for the most frequent 2 word-clusters in the two datasets.

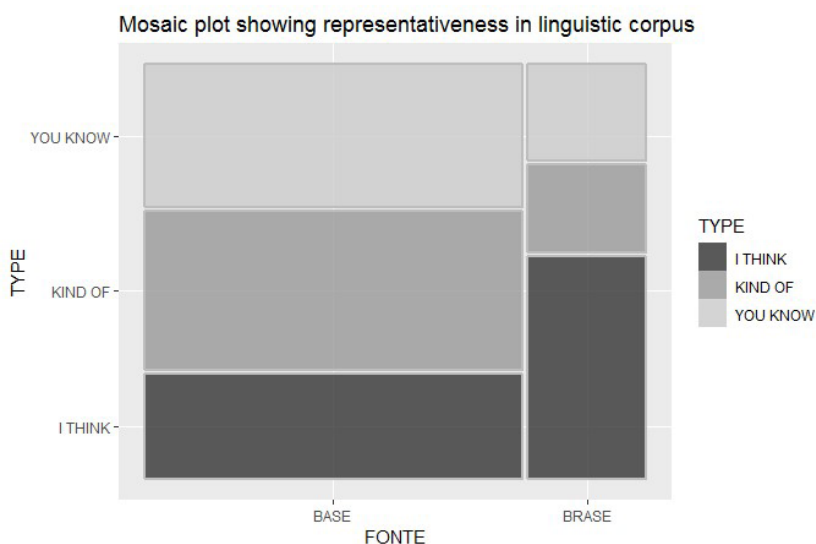
Table 3. Multiword units functioning as metalinguistic monitors in both corpora

BASE 150.000 words			BRASE 50.000 words		
Item	Raw freq.	Normalised Per 100.000	Item Freq. bruta	Raw freq.	Normalised Per 100.000
KIND OF A KIND OF	325	216	I THINK	143	286
YOU KNOW	293	195	YOU KNOW	62	124
I THINK	215	143	KIND OF	56	112
SORT OF	212	141	A LITTLE	25	50
I MEAN	73	48			
A SENSE	50				
TOTALS	2.340	861	TOTALS	286	572

Following Cortes (2002), a cut-off point of 20 occurrences per 100,000 words was established for the analysis of the multiwords units. A brief view of the list generated reveals interesting differences between the two datasets. The number of clusters in BASE outnumbers BRASE and the frequency in BASE is also higher than it is in the Brazilian data.

### 5.3. Statistical Analysis for 2-word clusters

When the use of multiword units is regarded, it is possible to verify, through the analysis of Graph 2, that there is a greater frequency of the expression *I think* in BRASE and of *kind of* in BASE, it is also possible to identify in the graph the size difference of the two corpora. That said, the frequency presented in the samples of these terms is displayed in a more homogeneous form in each corpus than it was in relation to the frequency of the words as shown in Graph 1.



Graph 2. Mosaic display for sample of multiword units in the BASE and BRASE corpora

With respect to the results found in the analysis of the contingency tables involving the multiword units displayed in Table 3, we can observe,

through the RR, that only the word combination *I think* presented a greater chance of occurring in BRASE in relation to BASE. Additionally, the same happened with respect to the result obtained from OR, that is, the word combination *I think* has a greater chance of being used in BRASE than in BASE (almost double the chance).

For the results obtained in comparison between the equality of the proportions of use for the words in BRASE and in BASE, in the four statistical measures found, p-value for chi-square test, p-value for equality of proportions and Confidence Interval confirmed that, at the level of 5% significance, the proportions differed. Likewise, the log-likelihood (LL) results were obtained, all of them were superior to the cut-off point of 3.8, associated with the significance level of 5%. These results confirm the overuse of *I think* in BRASE in relation to BASE, together with the underuse of the combinations *kind of* and *you know* in BRASE, when compared to BASE, as one can see in Table 4:

**Table 4. Statistical Measures of occurrence and association of multiword terms**

ITENS	RR	OR	P-VALUE $\chi^2$	LL	P-VALUE = PROP	IC (0,95) for D
I THINK	1,598	1,995	0	38,47	0	[0,099; 0,200]
YOU KNOW	0,876	0,635	0,001	-11,62	0,001	[-0,075; -0,036]
KIND OF	0,588	0,517	0	-24,17	0	[-0,139;- 0,067]

The next sections will deal with the items that demonstrated a statistical significance for the purpose of this study.

#### 5.4. The cluster *I think*: a sample of overuse

Holmes (1985; 1990) identifies two broad semantic categories of *I think*: deliberative and tentative. The former, according to Holmes (1985, p. 33), is used to express personal confidence in the proposition asserted and therefore adds weight to the speech act. The latter, is used to express uncertainty. In Table 4, we focused on the tentative function since the main aim of this paper is to concentrate on items functioning as face-saving devices and as negative politeness strategies (Brown & Levinson 1987; Goffman 1967). In

part, the overuse of *I think* can be attributed to first language interference. The expression *Eu acho* (Marcuschi 1989) in Portuguese is quite common when one is expressing his/her opinion about some issue.

However, it is important to state that the cluster *I think* might not be the most appropriate cluster to be used in an academic context as it is commonly associated with casual conversation (see McCarthy & Carter 2006). At this point of the analysis, it can be argued that this group of students are misusing the item *I think* in academic oral presentations. The examples show that they are borrowing an item from casual conversation and incorporating in their academic oral production in an excessive way.

(3) *I think that it might be a kind of revenge a kind of revenge. Revenge against the society against the society against the maybe against something that he has lift in the air ...*

Overall, the overuse of *I think* in learner's production may lead to an ineffective strategy of image projection and face-saving strategy, as it barely meets the interlocutor's expectations for exchanges held in the academic sphere.

In the British data, it can be noticed that the expression *I think* is commonly used clustering with other modal items reinforcing the preference for epistemic forms by native or near-native speakers interacting in an academic environment. In sum, our findings demonstrate that the use of *I think* is multifunctional. On the one hand, it displays uncertainty, while it also serves as an epistemic<sup>5</sup> marker, influenced by the learners' mother tongue (Marcuschi, 1989).

### 5.5. The cluster *you know*: a sample of underuse

According to Östman (1981) and Holmes (1986), *you know* serves a variety of different, though closely related functions in discourse. Particularly important for this study are Holmes' (1986) observations on the marker, in particular, as a hedge device. She divides the functions of *you know* into two categories: Category I comprises instances of *you know* expressing speaker confidence or certainty (positive politeness) and category II involves the usage of *you know* to express uncertainty of various kinds (negative politeness). Once again, only examples of *you know* functioning as a hedge

5 Epistemic modality refers to the degree of commitment one has in relation to what one says.

were isolated for analysis. *You know* occurs 124 times in BRASE, and 195 in BASE (normalised occurrences).

The Log-likelihood test indicates that within a hedging function, learners underuse the item when compared to native speakers. A thorough analysis of the concordance lines for *you know* in BRASE demonstrates that the item is more commonly used as a marker of assertiveness. In the majority of the examples, learners were expressing their certainty on a proposition whereas in BASE the opposite is observed. Extract (4) below illustrates the discussion carried out in this section.

(4) BRASE IFA1-B1<sup>6</sup>

*now you've said something . you said something interesting because when Lula was campaigning for the first time there was the same euphoria . about him **you know** and remember when Lula was elected for president the parties around the country it was . in a way the same kind of feeling*

In the case of *you know*, it can be said that learners rely more often on the function of assertiveness and/or shared knowledge than on the role of a hedge. This finding reinforces the claim that students are not aware of the items that are more commonly associated with the academic domain, in the case here academic oral presentations. The underuse of the cluster *you know* shows that learners usually lack the linguistic repertoire needed to cope with negative politeness strategies. As a result, they tend to express themselves in an assertive way, which can be interpreted as an imposition, not complying with the norms of academic interaction.

## 6. Final remarks

This paper focused on the function and meaning of the metalinguistic monitors: *just*, *maybe*, *I think*, *you know* and *kind of* in two corpora: the Brazilian Academic Spoken English (BRASE) Corpus and the British Academic Spoken English (BASE) Corpus. These items were analysed according to their occurrence in the main corpus (BRASE) and then compared and contrasted with the reference corpus (BASE). The analysis was

6 IFA 1 stands for *Inglês para Fins Acadêmicos 1 – English for Academic Purposes Level 1*. The example sentence was produced by students taking this course. B1 corresponds to their proficiency level (intermediate) following the *Common European Framework of Reference for Languages* (CEFR).

organised following the results from the statistical tests RR, OR, P-value and Log-likelihood. These tests were instrumental in identifying the features that were responsible for the linguistic differences between the two corpora, indicating the items worth being analysed in more detail.

Hence, identifying the overuse and/or underuse of items in students' spoken interlanguage may provide language teachers a better account of learners' production, in the case of this study, in oral academic discourse. Once teachers understand learners' discourse in a more accurate way, they can design activities that will better suit their needs.

In this study, learners used *just* employing its hedging function. The overuse of *just* in BRASE confirms that although learners seem to be aware of face issues, their choice of items differ from that of BASE speakers. Its high frequency of occurrence in BRASE indicates that learners are employing a marker from informal oral discourse in their oral academic presentations. This fact suggests that this group of Brazilian learners is unaware of the norms guiding academic context and thus need to develop their oral academic literacy in English.

As for the items *I think* and *you know*, the first point worth mentioning is the total number of clusters in both corpora. There are more clusters in BASE than in BRASE. In addition, the overall occurrence also points to significant differences. There is more variation in BASE: different forms are evenly distributed in the British corpus, while in BRASE the occurrences are concentrated on a more restricted set of forms. These findings suggest implications to the discourse produced by learners. By relying on a limited set of clusters, learners are constrained to a rigid repertoire, which limits their discourse and their ability to communicate in a more effective way, especially in an academic environment. Failing to use these markers can hinder communication and pose difficulties for students willing to participate in a more globalized academic community.

Focusing on linguistic aspects like the ones presented in this paper, we can possibly contribute to improve learners' oral communicative skills. In addition, a refined description of learners' spoken interlanguage contributes to raise cultural and linguistic awareness, guiding material design and to the growing area of Brazilian curriculum concerning English for Academic Purposes.

Overall, the prevailing deployment of less pragmatically enriched pragmatic markers in the learners' production analysed here may reveal that explicit instruction is a crucial element to raise awareness of these items in learner's discourse. As a consequence, this guidance would help pave the



way for more effective interpersonal interactions in the academic scenario, particularly with regard to politeness strategies and face-work.

Finally, it is worth mentioning that the statistical methods employed in this paper allowed for the achievement of an optimal point of data cross-analysis, which we believe has led to more reliable and significant results. For this reason, we believe that these findings may, and should, be replicated and expanded to other populations and to different scenarios.

## References

- Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
- Aijmer, K. (2002). *English discourse particles: Evidence from a corpus* (Vol. 10). Amsterdam: John Benjamins Publishing.
- Aijmer, K., & Simon-Vandenberg, A. M. (2004). A model and a methodology for the study of pragmatic markers: The semantic field of expectation. *Journal of Pragmatics*, 36(10), 1781–1805.
- Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies*, 3(1), 173–190.
- Aijmer, K. (2013). *Understanding pragmatic markers*. Edinburgh: Edinburgh University Press.
- Andersen, G. (2001). *Pragmatic markers and sociolinguistic variation: A relevance-theoretic approach to the language of adolescents* (Vol. 84). Amsterdam: John Benjamins.
- Biber, D., Johansson, S., Leech, G., & Conrad, S. E. Finegan (1999). *Longman Grammar of Spoken and Written English*. London: Longman.
- Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2), 97–116.
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge: Cambridge University Press.
- Brown, P. (2015). Politeness and language. In *The International Encyclopedia of the Social and Behavioural Sciences (IESBS)* (pp. 326–330). (2<sup>nd</sup> ed.) Elsevier.
- Bublitz, W. (1978). *Ausdrucksweisen der Sprechereinstellung im Deutschen und im Englischen*. Tübingen: Niemeyer.
- Cunha, G. X. (2015). As relações retóricas e a negociação de faces em debate eleitoral. *Confluência*, 1(47), 205–238.
- Cortes, V. (2002). Lexical bundles in freshman composition. *Using corpora to explore linguistic variation*, 9, 131–145.
- Downing, A., & Locke, P. (2006). *English grammar: A university course*. New York: Routledge.

- Erman, B. (2001). Pragmatic markers revisited with a focus on you know in adult and adolescent talk. *Journal of pragmatics*, 33(9), 1337–1359.
- Fraser, B. (1999). What are discourse markers? *Journal of pragmatics*, 31(7), 931–952.
- Fraser, B. (1990). An approach to discourse markers. *Journal of pragmatics*, 14(3), 383–398.
- Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied linguistics*, 28(3), 410–439.
- Goffman, E. (1967). *Interaction ritual: essays on face-to-face interaction*. New York: Anchor Books.
- Goffman, E. (1976). Replies and responses. *Language in society*, 5(3), 257–313.
- Halliday, M. A. K., Matthiessen, C., & Halliday, M. (2014). *An introduction to functional grammar*. London: Routledge.
- Holmes, J. (1985). Sex differences and miscommunication: Some data from New Zealand. *Cross-cultural Encounters: Communication and Miscommunication*, (pp. 24–43). Melbourne: River Seine.
- Holmes, J. (1986). Functions of you know in women's and men's speech. *Language in society*, 15(1), 1–21.
- Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language & Communication*, 10(3), 185–205.
- Haugh, M. (2013). Disentangling face, facework and in/politeness. *Pragmática sociocultural*, 1(1), 46.
- Kerbrat-Orecchioni, C. (2006). *Análise da conversação: Princípios e métodos*. São Paulo: Parábola Editorial.
- Kriwonossow, A. (1977). *Die modalen Partikeln in der deutschen Gegenwartssprache*. Vol. GAG 214. Göppingen: Kümmerle-Verlag.
- Marcuschi, L. A. (1989). *Marcadores conversacionais do português brasileiro: formas, posições e funções*. Campinas: Editora da UNICAMP.
- McCarthy, M. & Carter, R. (2006). This that and the other: *Multi-word clusters in spoken English* as visible patterns of interaction. *Explorations in corpus linguistics*, 7–26.
- O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
- Oliveira, A. L. Adorno Marciotto, G. X. Cunha, & M. Vieira Miranda (2017). Nominalizations as complex strategies of politeness and face-work in scientific papers written in Brazilian Portuguese. *Cadernos de Estudos Lingüísticos*, 59(2), 361–374.
- Orfanò, B.M. (2010). *The representation of spoken language: a corpus-based study of sitcom discourse* [unpublished PhD dissertation], Limerick: Mary Immaculate 2College-University.
- Östman, J. O. (1981). *'You Know': A discourse-functional study*. Amsterdam: John Benjamins.
- Ran, Y. (2003). A pragmatic account of the discourse marker WELL. *Journal of Foreign Languages*, 3, 58–64.

Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora* (pp. 1–6) [held in conjunction with the 38<sup>th</sup> annual meeting of the Association for Computational Linguistics (ACL 2000). 1–8 October 2000], Hong Kong.

[submitted on March 31, 2018 and accepted for publication on September 17, 2018]

# CORPUS STYLISTICS IN TRANSLATION-ORIENTED TEXT ANALYSIS: APPROACHING THE WORK OF DENTON WELCH FROM A FUNCTIONALIST PERSPECTIVE

## ESTILÍSTICA DE CORPUS E ANÁLISE TEXTUAL DE RELEVÂNCIA TRADUTÓRIA: UMA ABORDAGEM INICIAL À OBRA DE DENTON WELCH A PARTIR DE UMA PERSPECTIVA FUNCIONALISTA

Guilherme da Silva Braga\*  
guizomail@gmail.com

This article is an effort towards interpreting the findings of a corpus-based stylistic analysis of the short narrative “Sickert at St Peter’s” (1942), written by the English writer and painter Denton Welch (1915–1948), within the larger framework for translation-oriented text analysis presented by Christiane Nord in *Textanalyse und Übersetzen* (2009). The aim is to explore both the theoretical possibilities and the practical applications of a corpus-based approach to the lexical analysis phase of Nord’s model from a literary translation perspective, in which style and word choice play a critical role. Once the statistical findings of the corpus-based analysis are presented, the 25 highest-ranking keywords in the text are analyzed in context. Translation briefs and literary translation in general are discussed, and a global pre-translational strategy for translating “Sickert at St Peter’s” is presented. By way of conclusion, it is argued that the method described promotes valuable insights for literary interpretation and serves as a practical aid in developing a pre-translational strategy for translating individual texts.

**Keywords:** Corpus Linguistics. Corpus Stylistics. Literary Translation. Translation Studies. Functionalism. Denton Welch.

O presente artigo propõe-se a interpretar os resultados de uma análise estilística da narrativa breve “Sickert at St Peter’s” (1942), do escritor e pintor inglês Denton Welch (1915–1948), feita de acordo com os métodos da linguística de corpus no contexto da análise textual de relevância tradutória apresentada por Christiane Nord em *Textanalyse und Übersetzen* (2009). O objetivo é explorar as possibilidades

---

\* Universidade de Coimbra, Portugal.

teóricas e as aplicações práticas de uma abordagem de corpus na fase de análise lexical que compõe o modelo de Nord sob a perspectiva da tradução literária, em que o estilo e as escolhas lexicais desempenham um papel fundamental. Após uma apresentação dos resultados da análise, as 25 palavras-chave de maior destaque são analisadas em contexto. Procede-se então a uma discussão sobre as especificações tradutórias e a tradução literária em geral e à apresentação de uma estratégia pré-tradutória global passível de ser aplicada à tradução de “Sickert at St Peter’s”. O estudo conclui que a metodologia descrita oferece um contributo valioso para a interpretação de textos literários e serve como uma ferramenta prática no desenvolvimento de estratégias pré-tradutórias aplicáveis a textos literários individualmente considerados.

**Palavras-chave:** Linguística de corpus. Estilística de corpus. Tradução literária. Estudos de tradução. Funcionalismo. Denton Welch.



## 1. Translation-oriented text analysis: An overview

In her book *Textanalyse und Übersetzen* (2009), Christiane Nord proposes a model for translation-oriented text analysis that aims to produce target texts characterized as *functional* – that is, suitable to the Skopos (purpose) to be achieved in the target culture. This functionalist approach recognizes that different translation purposes call for different translation approaches and claims that a thorough analysis of source text features offers valuable insights as to how a given source text works – and by extension as to which should be the optimal translation procedure in each particular case in view of these findings and the translation brief for the translation to be carried out.

According to Nord’s model, any text can be analyzed within a framework composed by seventeen items. Eight of these are extratextual elements (“sender”, “sender’s intention”, “audience”, “medium/channel”, “place of communication”, “time of communication”, “motive for communication” and “text function”) and eight are intratextual elements (“subject matter”, “content”, “presuppositions”, “text composition”, “non-verbal elements”, “lexis”, “sentence structure” and “suprasegmental features”). Together, these sixteen

textual elements work in order to achieve a given “effect”<sup>1</sup> – at the same time, the seventeenth item to be analyzed and a special category defined as “ein übergreifender Faktor, durch den das Zusammen’spiel zwischen textexternen und textinternen Faktoren erfasst wird” (Nord 2009, p. 40).<sup>2</sup>

The last three intratextual items in the exhaustive list presented above are described by Nord as the “sprachlich-stylistische Merkmale” (Nord 2009, pp. 89-90)<sup>3</sup> of a text, and as such can be considered as particularly relevant in the translation of literature, given that style and aesthetics have always played a major role in the production and reception of literary texts.<sup>4</sup> The very notion of literary language, regardless of a precise definition of what constitutes literature<sup>5</sup>, seems to be substantially dependent on style:

Whatever stand we take on these questions of definition, literary language is clearly assumed to have a particularly connotative, expressive or aesthetic meaning of its own. (Nord 1997, p. 81)

This article is an effort to explore ways of employing a corpus linguistics approach in the pre-translational stage of a literary translation as part of the broader translation-oriented text analysis framework set forth by Nord.

Once the theoretical possibilities of this approach are discussed, its practical application is demonstrated by means of a case study centered around the short narrative “Sickert at St Peter’s” by the English writer and painter Maurice Denton Welch (1915–1948). After a corpus-based lexical analysis of this particular narrative is made against the background of Welch’s three full-length novels, the results of the corpus analysis are examined in context, a translation brief for a literary translation of the narrative is defined and a pre-translational global translation strategy is presented.

---

1 The English-language terminology, as well as all English translations of Nord’s original German text, are henceforth taken without exception from *Text Analysis in Translation*, translated into English by Christiane Nord and Penelope Sparrow.

2 “A global or holistic concept, which comprises the interdependence or interplay of extratextual and intratextual factors” (Nord 2009, p. 42).

3 “Formal-aesthetic characteristics” (Nord 2005, p. 91).

4 See Herrmann, Dalen-Oskam and Schöch (2015) for a comprehensive review of definitions of style at different periods and in different cultures.

5 Nord defines literariness as “first and foremost a pragmatic quality assigned to a particular text in the communicative situation by its users” (Nord 1997, p. 82).

## 2. Integrating a corpus approach to translation-oriented text analysis

The integration of a corpus approach into the larger framework in which Nord's functionalist model for translation-oriented text analysis works does not present any sort of theoretical restriction, since functionalism explicitly "makes use of descriptive methods" (Nord 1997, p. 2) such as corpus linguistics. When applied specifically to the analysis of literary texts, the corpus linguistics approach has often been called "corpus stylistics".

Several studies of literary language from a corpus stylistics perspective have made interesting contributions to the field<sup>6</sup> by focusing on the computer-assisted generation of a keyword list for the text to be analyzed and a human-made interpretative analysis of the highest-ranked keywords in this keyword list. The initial automated phase is often called quantitative analysis, whereas the later interpretative phase receives the name of qualitative analysis.

Keywords can be defined as words whose frequency in the text under analysis is statistically significant when compared to the frequency of those same words in a reference corpus formed by any number of other texts. In other words, given a text *T* and a reference corpus *RC*, a keyword list of *T* is a list of words whose frequency in *T* is proportionally higher than would be expected from the frequency observed in *RC*. During the quantitative analysis phase, there are several different methods to make these calculations (some of which are built into computer programs specifically designed for corpus analysis), but the result is always a numerical keyness value ascribed to each of the words in *T*. The higher the keyness, the larger the deviation between the expected word frequency based on the statistical data provided by *RC* and the actual data measured in *T*. Hence, the top-ranking words in a computer-generated keyword list can and should be treated as likely candidates for further qualitative analysis.

This kind of approach can be applied to Nord's unmodified model for translation-oriented text analysis<sup>7</sup>, even though the application of a corpus approach to literary texts presents special problems.

The basic text typology espoused by functionalism relies on three (Vermeer and Reiß 1991) or four (Nord 2014) basic textual functions: the

---

6 See the bibliography for the examples cited in this article.

7 See Nord (2009), pp. 124–131 or Nord (2005), pp. 122–129.

informative/referential function, the expressive function, the operative/appellative function and the phatic function (the last being exclusive to Nord). However, this classification serves only as a description of the manner in which the text is intended to work. The subject matter and the formal characteristics of texts are encompassed by the notion of text classes<sup>8</sup>, defined as texts “classified according to linguistic characteristics of conventions” (Nord 1997, p. 37).

Several text classes<sup>9</sup> have a more or less predictable form in a given culture – this holds true both in the case of highly codified and topic-specific types with very strict form, such as cooking recipes and weather reports, as well as topic-independent but form-bound types, such as newspaper items. Any text which does not conform to the expected standard can be said to be a deviation (in a neutral sense) from the given text class’s norm.

Literary texts, however, seem to pose a special problem to the notion of text classes: though texts can in fact be grouped under the description of “poems”, “short stories” or “novels”, these do not have a standard form, a standard theme or even a standard style of writing, all of which depend entirely on the particular artistic project of each individual author.<sup>10</sup> As a result, there is absolutely no standard for a piece of literary writing, which can – both in theory and practice – be written in any style, devoted to any topic and as short or as long as the individual author wishes. As pointed out by Nord (2009):

Im Bereich der literarischen Texte sind konventionelle Merkmale nicht so häufig wie bei den Gebrauchstexten. Gattungsbezeichnungen wie Roman, Kurzgeschichte, Anekdote weisen zwar darauf hin, dass man von den so klassifizierten Texten bestimmte gemeinsame Merkmale erwartet, diese beziehen sich aber meist auf inhaltlich-thematische (z.B. Anekdote vs. Witz), extensionale (z.B. Roman vs. Erzählung) oder epochenspezifische (z.B. Novelle vs. Kurzgeschichte) Aspekte oder bestimmte Stileigenschaften (z.B. *Sturm und Drang*). Im Allgemeinen wird jedoch der literarische Einzeltext als Ergebnis

8 Nord’s terminology is unstable with relation to this term, which corresponds to the German *Textsorte*: in Nord (2005), the term has been rendered in English as “text class” (see p. 20), whereas Nord (1997) refers to the same concept as “text genres or varieties” (p. 37).

9 For a thorough discussion of text types (*Texttypen*) and text classes (*Textsorten*), see Nord (2009), pp. 19–21, or Nord (2005), also pp. 19–21.

10 Nord (2009, p. 46; 2005, pp. 47–48) makes a clear distinction between text sender and text producer. The sender is responsible for the communication being carried out, whereas the producer is responsible for producing the text which is to serve as a means of communication. The designation of “author” is employed only when the roles of sender and producer coincide in one and the same person.



eines individuellen Schöpfungsprozesses gesehen, der gerade dadurch seine (künstlerische) Bedeutung erhält, dass er *nicht* vorhandene Muster reproduziert (...), sondern "originell" und damit innovatorisch ist.<sup>11</sup>

The situation described above is not without implications for corpus analyses of literary works, since this kind of approach often relies on reference corpora whose purpose is to serve as a balanced and neutral reference against which the unique features of the corpus being analyzed may be revealed – but, given that not even literary texts within the same text class present a minimally standardized style or a minimally established theme, it is simply not possible to find a balanced and neutral reference corpus to serve as a means of comparison. As a result, the choice of adequate reference corpora for analyzing a literary corpus needs to be made with due attention to the specifics of the task at hand.

In order to cope with the problem described above, scholars involved in corpus stylistics have largely resorted to two different work methods, and often to a combination of the two.

The first method consists in comparing a corpus of works by the author to be studied against a corpus of various works by several contemporary authors. This procedure may be used to prevent the resulting keyword list from including words that are not peculiar to the author in question, but rather common in all literature written at that particular time. Patrick Maiwald (2011) offers a clear illustration of the problem: when comparing George MacDonald's (1824–1905) "fantasy" works to the "imaginative" subset of the 20<sup>th</sup> century British National Corpus, words such as "light", "shine", "ray", "gleam", "glimmer", "moonlight" and "sunlight" appear as keywords – but, in a comparison of MacDonald's fantasy works with the work of other 19<sup>th</sup> century writers, they disappear completely from the keyword list, "thus proving that this preoccupation with light and 'visuality' is not particular to MacDonald, but to Victorian writers in general" (Maiwald 2011, p. 73). Jonathan Culpeper offers a similar warning:

---

11 (Nord 2009, p. 21). The English translation reads: "In the field of literary texts conventional elements are not so frequent as in the field of non-literary texts. Designations such as "novel", "short story", or "anecdote" may, however, indicate that the texts belonging to one of these genres are expected to possess certain common features. Literary genres are often differentiated by special features of subject matter or content (anecdote vs. joke), extension (novel vs. short story) or by their affiliation to a literary era (novella vs. short story), as well as by certain stylistic properties. Nevertheless, a literary text usually has to be regarded as the result of an individual and creative process. Its (artistic) significance lies precisely in the fact that it does *not* reproduce existing text models (...), but represents an original innovation" (Nord 2005, pp. 21–22).

In any keyword analysis, the choice of data for comparison (the reference list) is crucial. (...) Clearly, a set of data which has no relationship with the data to be examined is unlikely to reveal interesting results. (Culpeper 2002, p. 15)

When judiciously applied, though, this method offers valuable insights into the general style and themes of a given author.

The second method consists in using a corpus of works by a single author and comparing the individual work to be studied against some of all of the others. This is a rather more specific method than the previous one in that it will not offer insights related to the general writing traits of the author in question, but rather outline what makes the single work – or even component parts of the single work, like chapters or the speech of a single character – unique. This method may lead to unexpected finds that could hardly be gleaned from a stylistic analysis performed without any sort of computer assistance: after preparing separate corpora for six individual characters in Shakespeare's *Romeo and Juliet*, for example, Jonathan Culpeper found that Juliet's most prevalent keyword is "if" – a find described as "striking" because "it does not seem so obviously guessable, partly because it is a grammatical word" (Culpeper 2002, p. 20).

Once a fitting corpus stylistics method is chosen and implemented, the analysis of the resulting keyword list as part of the lexical analysis phase within the larger framework for translation-oriented text analysis can proceed.

### **3. Case study: A corpus stylistics analysis of Denton Welch's "Sickert at St Peter's"**

In order to illustrate the approach outlined above, I would like to present a corpus-based, translation-oriented pre-translational text analysis of the short narrative "Sickert at St Peter's" by the English writer and painter Maurice Denton Welch (1915–1948).

Welch was an art student and an aspiring painter when, at the age of 20, he was hit by a car while riding his bicycle, suffering devastating injuries as a result. His spine was fractured; his kidneys failed; his bladder was paralyzed and he was left partially impotent. However, in spite of an impressive recovery which eventually allowed him to resume walking and even riding a bicycle, for the rest of his life Welch had to deal with severe long-term injuries caused by the accident.

After making a partial recovery, Welch left the hospital for a nursing home, and while living there paid a visit to the English painter Walter Sickert (1860–1942) in 1936. Years later, as Welch became more interested in writing – partly as a kind of post-accident personal therapy –, this episode found a written form and eventually resulted in his first published piece: the short narrative “Sickert at St Peter’s”, published in the August 1942 issue of the literary magazine *Horizon*. After this first publication, Welch would go on to write poems, journals, around sixty short stories and the three full-length novels for which he is most known, entitled *Maiden Voyage* (1943), *In Youth Is Pleasure* (1944) and the posthumous *A Voice Through a Cloud* (1950). This last work describes in great detail the accident whose consequences finally claimed Welch’s life at 33.

It should be noted here that the context provided above is not presented as a mere curiosity; as previously stated, Nord’s model for translation-oriented text analysis includes several extratextual items which should be subjected to scrutiny, and among these are “sender”, “medium/channel”, “place of communication”, “time of communication” and “motive for communication”, all of which are here accounted for, with the possible exception of “sender”.

Literary pieces have an “implicit narrator” which should not be confused with the author.<sup>12</sup> However, as Welch scholar and biographer Michael De-la-Noy notes in several occasions, the works of Denton Welch – described as “an exclusively autobiographical author” (De-la-Noy 1984, p. xi) – tread a very fine line between fiction and autobiography:

(...) the line between fact and fiction in [Welch’s] work is often as narrow as any writer could have drawn it. (De-la-Noy 1987, p. 9)

Although [Welch] occasionally juggled with events for dramatic purposes, every occasion about which he wrote and every character he wrote about was taken from real life. (De-la-Noy 1984, p. viii)

(...) no writer has mirrored his life in his work so transparently, nor left us such poignant evidence of this integral connection. (De-la-Noy 1987, p. 10)

---

12 As Nord puts it: “In fiktionalen Texten wird (...) ein ‘impliziter Erzähler’ eingeführt, der nicht mit dem Autor gleichzusetzen ist” (Nord 2009, p. 126). The English translation reads: “In fictional texts, we have to assume an ‘implicit narrator’ who is not identical with the author” (Nord 2005, p. 124).

In view of the above, it remains unclear to what extent Welch's first-person narrator in "Sickert at St Peter's" should be regarded as the 'same' person as the author of the text, even though this observation has little impact on the lexical analysis being proposed.

### 3.1. Method

Initially, Welch's three novels were loaded one by one into the corpus analysis software AntConc 3.4.4w and processed to produce an individual, lemmatized<sup>13</sup> word list for each novel by using a slightly modified version of Yasumasa Someya's "lemma\_no\_hyphen.txt" lemma list<sup>14</sup> for English. In each case, the resulting word list was saved as a separate text file. Next, the short narrative "Sickert at St Peter's" was loaded in AntConc, lemmatized with Someya's lemma list and, with the previously created lemmatized word lists for the three novels set as reference corpora, processed for both positive and negative keyword lists using the log-likelihood method. The resulting keyword list provides data about words which appear unusually often in "Sickert at St Peter's" in relation to the Welch's three full-length novels and about words which, given their frequency in the novels, would also be expected to appear in the short narrative, but do not do so (or do so at a significantly lower rate).

The purpose of this keyword list is to identify what makes "Sickert at St Peter's" different from Welch's novels in terms of constituent lexical items – a procedure which should promote valuable insights into the piece as well as serve as a practical aid in developing a pre-translational strategy for translating this particular text.

---

13 Lemmatization is the process by which all the different forms of a given word (plurals, conjugated verbs etc.) are interpreted by the software as being one and the same.

14 A lemma list is a computer-readable series of dictionary-entry word forms and other different forms (regular and irregular plurals, regular and irregular verb conjugations etc.) these words may assume. As long as an inflected form is associated with its corresponding dictionary-entry form in a lemma list, specialized software will recognize the inflected form as being the same as the dictionary-entry form of the word in question. In the case of Yasumasa Someya's "lemma\_no\_hyphen.txt" lemma list for English, the modification consisted simply in removing two-letter and three-letter acronyms with an apostrophized plural or past form from the lemma list (e.g.: "WC's" as the plural of "WC" and "KO'd" as the past of "KO"). Since by default apostrophes are not recognized as characters by AntConc, the presence of these apostrophized forms in the lemma list caused the software to register <s> and <d> as forms of "WC" and "KO", respectively. No other changes were made.

### 3.2. Quantitative Analysis of Keyword List Items

The resulting keyword list for “Sickert at St Peter’s”, up to the 25<sup>th</sup> keyword, is as follows:

**Table 1. Resulting keyword list for “Sickert at St Peter’s”**

Rank	Count	Keyness	Keyword
1	38	346.594	sickert
2	9	82.088	raven
3	8	72.967	gerald
4	7	43.287	boot
5	7	35.915	photograph
6	11	33.545	mrs
7	3	22.885	eden
8	2	18.242	anthony
9	2	18.242	beaverbrook
10	9	16.962	picture
11	12	14.800	us
12	2	14.444	hearth
13	2	14.444	original
14	2	14.444	sewer
15	18	13.277	t
16	4	13.210	evidently
17	2	12.739	pit
18	4	11.829	art
19	3	11.813	grunt
20	2	11.575	stringy
21	3	10.516	accident
22	5	10.337	cup
23	2	9.971	famous
24	3	9.719	sofa
25	13	9.226	room

In order to interpret the keyness value of each word in the list, it is necessary to understand how high the keyness value has to be in order to be considered statistically significant. Standard values are presented in the table below<sup>15</sup>:

**Table 2. Significant values for keyness analysis**

Critical value	Percentile	Error margin
3.84	95%	5%
6.63	99%	1%
10.83	99.9%	0.1%
15.13	99.99%	0.01%

In practice, this means that a word identified as having a keyness of [critical value] has a [percentile] percent chance of being statistically significant, and an [error margin] percent chance of being purely accidental.

As the keyness values in the keyword list for “Sickert at St Peter’s” show, all keywords ranked 1 to 25 fall either in the very low 1% (words ranked 21 to 25) or in the even lower 0.1% (words ranked 1 to 20) error margin. Once this has been ascertained, a qualitative analysis of the keywords in the list can be undertaken.

### 3.3. Qualitative Analysis of Keyword List Items

The keyword list generated during the quantitative analysis evidences potentially rich aspects of the text to be analyzed, but the compiled results should be considered only as the starting point for a full-scale corpus stylistics investigation. As Michaela Mahlberg points out, “quantitative research can only provide valuable insights when it is linked to qualitative analysis” (Mahlberg 2010, p. 292). This claim is endorsed by Dan McIntyre and Brian Walker, who wrote:

Of course, key comparisons can only be a starting point. In order to fully understand the lists produced by a computer tool, we must return to the text. Quantitative analysis guides qualitative analysis, which might guide further quantitative analysis. (McIntyre & Walker 2010, p. 522)

15 Adapted from “Log-likelihood and effect size calculator”, available at <http://ucrel.lancs.ac.uk/llwizard.html>.

Let us then return to the text: on the top of the keyword list we find the name “Sickert”, which also appears as the first word in the title as the explicit theme of the narrative and whose importance is therefore apparent.<sup>16</sup> This is not a surprising find at all, even though it should be observed that in many cases “Sickert” refers to Sickert’s wife, “Mrs. Sickert” – the only female character in the story, but also undoubtedly an important figure, as the one and only referent to all 11 occurrences of “mrs” (keyness rank: 6).

Apart from the proper nouns “Raven” (keyness rank: 2), “Gerald” (keyness rank: 3), “Anthony Eden” (keyness rank: 8 for “Anthony” and 7 for “Eden”) and “Beaverbrook” (keyness rank: 9), whose presence in the keyword list can be accounted for by the simple reason that these are the names of the characters of the story, there is a noticeably high proportion of common nouns related to concrete objects: “boot” (keyness rank: 4), “photograph” (keyness rank: 5), “picture” (keyness rank: 10), “hearth” (keyness rank: 12), “sewer” (keyness rank: 14), “pit” (keyness rank: 17), “sofa” (keyness rank: 24) and “room” (keyness rank: 24) account for a full one-third of the 25 most relevant keywords. This find seems to corroborate the textual impression of Welch’s fascination with physical objects, which can be seen in the following examples (henceforth, all keywords in the excerpts quoted shall be underlined):

My cup was of that white china which is decorated with a gold trefoil in the centre of each piece. Gerald’s was quite different. It was acid-blue, I think, with an unpleasant black handle and stripe; but I noted that both our spoons were flimsy and old. I turned mine over and saw, amongst the other hall-marks, the little head of George III winking up at me.

I looked at the other things on the table, at the brown enamel teapot, the familiar red and blue Huntley and Palmer’s tin, and at the strange loaf which seemed neither bread nor cake.

We discussed the various objects in the room. She told me that the two glittering monstrosities had come from a Russian church. We went up to them and I took one of the sparkling things in my hands. The blue and white paste lustres were backed with tinsel. They were fascinatingly gaudy and I coveted them.

---

16 While discussing text composition and text structure, Nord (2009, pp. 112–113) affirms: “Angesichts des besonderen Bedeutung von Textanfang und Textschluss für Verständnis und Interpretation eines Textes müssen bei der Analyse gerade diese Textteile besonders aufmerksam auf ihre rezeptions- und wirkungssteuernde Funktion hin untersucht werden”. The English translation (Nord 2005, p. 111) reads: “The special part that the beginning and the end of a text play in its comprehension and interpretation means that these may have to be analysed in detail in order to find out how they guide the reception process and influence the effect of the whole text”.

The example below illustrates the point even more poignantly: in a short excerpt which corresponds to a time-frame of probably less than one second, the narrator describes the characteristics of an object even before he understands what it is:

At last he brought out a rather crumpled, shiny object, and I saw that it was a photograph.

As the first common noun to appear in the keyword list, the word “boot” should be given due attention. It should also be noted that the noun “sewer” appears exclusively as a modifier to “boot” in the compound word “sewer-boot”, and that all these occurrences of “boot” refer to the incongruous boots that Sickert wore when he received the narrator and Gerald at home. The narrator is understandably taken by surprise: “from his toes to his thighs reached what I can only describe as sewer-boots”. The boots are also mentioned twice when Sickert performs a strange boisterous dance and explained during a scene in which Welch feels embarrassed because of a comment made by Sickert and casts down his eyes, which then rest on the painter’s boots. At this point, the text explicitly reads, “I was not thinking of his boots” – but Sickert notices Welch’s gaze and goes on to explain the reason for such an unusual piece of footwear while indoors:

‘Ah, I see that you’re staring at my boots! Do you know why I wear them? Well, I’ll tell you. Lord Beaverbrook asked me to a party and I was late, so I jumped into a taxi and said: “Drive as fast as you can!” Of course, we had an accident and I was thrown on to my knees and my legs were badly knocked about; so now I wear these as a protection.’

Attention should be here given to the word “accident” (keyness rank: 21), which appears in the opening single-sentence paragraph to “Sickert at St Peter’s”<sup>17</sup>:

I had been in Broadstairs for months, trying to recover some sort of health after a serious road accident.

As mentioned before, the accident in question had resulted in a broken spine which left Welch in bed for months, completely unable to walk. While he was at a nursing home, his doctor – knowing that he was an art student

---

17 See previous footnote.



– tried to persuade Sickert to pay him a visit, but Sickert would not hear about it. However, once Welch got back to his feet, Sickert agreed to receive him at home. If these two figures – until then unknown to each other, except for Welch’s previous acquaintance with Sickert’s works – already had a shared interest in art, at this point Welch discovers that Sickert has also suffered an accident that affected his legs. With this in mind, it becomes possible to argue that the boots are here used as a subjective identification device given a concrete form: just as Sickert’s “pictures” (keyness rank: 10) can be interpreted as a physical manifestation of a shared interest in “art” (keyness rank: 18), Sickert’s “boots” (keyness rank: 4) can likewise be interpreted as the physical manifestation of a shared fate in the form of an “accident” – a word which, in spite of occupying only the 21<sup>st</sup> place in the keyword list, could arguably be the object of a qualitative claim for greater relevance due to its presence in the opening sentence-paragraph of the narrative, as well as its direct relation to Welch’s life. This interpretation would underline the importance of Welch’s accident in the narrative, while at the same time using the accident theme – both on a conceptual and on a purely linguistic level – as an additional means for implying a bond with Sickert, whose eccentric footwear and behavior single him out as an artistic personality from the start. This interpretation seems to be corroborated at a later point in the narrative, when Sickert takes to his boisterous sewer-boots dance for the second time:

Then, as [Mr. Raven] passed Sickert on his way to the door, he felt in his pocket and with almost incredible courage brought out the crumpled little photograph again.

(...)

Sickert gave the same enigmatic grunt. It was somehow quite baffling and insulting.

Mr. Raven crept unhappily to the door and Mrs. Sickert followed swiftly to put salve on his wounds. Immediately Raven was out of the room Sickert became boisterous. He started to dance again, thumping his great boots on the floor. Gerald and I caught some of his gaiety. We did not mention Raven, but I knew that we were all celebrating his defeat. It was pleasant to feel that Sickert treated us as fellow artists. I wondered how many people each year asked him to paint pictures for love.

The passage above refers to Sickert’s final refusal of Mr. Raven’s tacitly made request for a free oil portrait of his mother, to be made based on the little “photograph” (keyness rank: 5) he produces. Each one of these

requests is rudely dismissed by Sickert with a “grunt” (keyness rank: 19), which – just as Sickert’s sewer-boots – can be read as a manifestation of Sickert’s unusual demeanor.

As it should be clear from the last quoted excerpt, the idea of a shared “us” (keyness rank: 11) identity as “fellow artists” (“art” being a word with a keyness rank of 18) is explicitly present in the text just alongside “pictures”, which in eight out of nine occurrences is employed as a synonym for “painting”. This is even more striking because, in spite of other characters like Mrs. Sickert and Mr. Raven being present and active in the story, all twelve instances of “us” as employed by the narrator refer exclusively to the characters involved with art – at the beginning of the story, the “art student” narrator and Gerald, his “art school friend”. But once Sickert appears in the narrative, he too joins the company described as “us”, while Mrs. Sickert moves around the house almost as if to leave the scene every time the pronoun is to be used. The narrator’s last comment, in which he expresses support for the rude dismissal of yet another painting request “for love”, has the additional effect of validating Sickert’s previously made apology for having refused a visit to the nursing home where the narrator was recovering from the accident:

I’m very sorry I didn’t come and see you, but I can’t go round visiting. (...) You see I have to keep painting all these pictures because I’m so poor.

Here, the high-keyness occurrences of the word “t” (keyness rank: 15), which upon closer inspection is revealed to be the final “t” in contractions of “not”, serves to illustrate the kind of discovery which could hardly be made without resource to a corpus approach. Such a high position in the keyword list unequivocally indicates that the tone adopted by Welch in “Sickert at St Peter’s” is considerably more informal than his three novels considered as a single corpus.

The presence of the word “evidently” (keyness rank: 16) in the keyword list would require additional research in order to be fully explained, but a preliminary study of the negative keywords in “Sickert at St Peter’s” shows the modal verbs “would” and “seem” at rank 3 (keyness value: 4.920) and 4 (keyness value: 3.750) respectively. When considered together, these observations would seem to imply that, whereas “Sickert at St Peter’s” is written in a quite direct and straightforward style, Welch’s later novels lean towards a more nuanced style. If confirmed, this find would constitute statistical evidence of Welch’s evolution as a writer.

The remaining eight words on the keyword list cannot always be clearly interpreted as particularly relevant in the context of “Sickert at St Peter’s”. “Hearth” is notably challenging from a qualitative point of view, since the two occurrences seem to be almost incidental, even though one of them is related to Sickert’s initial bout of boisterous dancing. At first sight, the two instances of “original” might suggest a relation with art, but when read in context it becomes clear that this is hardly the case: the first occurrence is employed as a synonym for “unmodified” in the cluster “original hall”, and the second actually refers rather counterintuitively to “an original Punch drawing” whose composition Sickert happened to be using in one of his paintings – in this case, “original” is used in the sense of “model”. “Pit” appears exclusively as a word used in the description of one of Sickert’s paintings (namely *The Miner*, though the painting is never mentioned by name), and might therefore be explained as an incidental textual item whose presence in the keyword list could be best explained by the specific mention to Sickert’s painting and its almost complete absence in Welch’s three other novels: the noun appears only once in the completely unrelated cluster “pit of [one’s] stomach” in *In Youth Is Pleasure*. “Stringy” appears only three times in Welch’s three novels (twice in *Maiden Voyage* in descriptions of people and once in *A Voice Through a Cloud* in the description of a coverlet), so that two occurrences in a text as short as “Sickert at St Peter’s” would appear to lend special significance to the term; but the character described as “stringy” is never presented by name, never says a single word to the narrator or his friend and leaves the house as soon as Sickert appears, never to return again. “Famous” may once again suggest a relation to art, but its two occurrences are rather unspecific: the first refers to the narrator’s attempted mental guess (“Perhaps she was someone famous”) – never confirmed or refuted – as to the identity of a woman in a photograph shown by Sickert; the second is used as a general *ad hoc* synonym for “politicians”. The remaining three words – “cup”, “sofa” and “room” – are likewise quite unremarkable individually, though “room” could probably be explored further as a means through which Sickert’s strong presence is insinuated in sentences such as “he waved his hand round the room”, “he shouted out in ringing tones for the whole room to hear” and “he called out across the room” – though it should be noted that some of the occurrences appear in unrelated compound nouns like “dining-room”, “drawing room” and “cloak-room”.

Having thus accounted for all 25 keywords in the list and offered an interpretation to each one of them in the context of “Sickert at St Peter’s”, I shall now turn to a discussion of the desired translation brief for “Sickert at St Peter’s”.

#### 4. Translation Brief

The general top-down approach proposed by Nord regarding translation decisions is as follows:

(...) a functional translation process should start on the pragmatic level by deciding on the intended function of the translation (documentary vs instrumental<sup>18</sup>). A distinction is then made between those functional elements of the source text that will have to be reproduced 'as such' and the ones that must be adapted to the addressee's background knowledge, expectations, and communicative needs or to such factors as medium restrictions and deixis requirements.

The translation type then determines whether the translated text should conform to source-culture or target-culture conventions with regard to translation style. (Nord 1997, p. 68)

Since functionalist approaches do not tell one how to translate, but only that one must translate according to the Skopos (purpose) to be achieved by the translation in the target-culture, there are no *a priori* conditions with regard to the characteristics of the target-text to be produced: these depend entirely on a set of parameters known as a translation brief (Übersetzungsauftrag), which "in an ideal case (...) would give as many details as possible about the purpose, explaining the addressees, time, place, occasion and medium of the intended communication and the function the text is intended to have" (Nord 1997, p. 30).

---

18 Nord explains the difference between the two basic types of translation – "documentary" and "instrumental": "The first aims at producing in the target language a kind of document of (certain aspects of) a communicative interaction in which a source-culture sender communicates with a source-culture audience via the source text under source-culture conditions. The second aims at producing in the target language an instrument for a new communicative interaction between the source-culture sender and a target-culture audience, using (certain aspects of) the source text as a model. (...) The result of a documentary translation process is a text whose main function is metatextual (...). The target text, in this case, is a text about a text, or about one or more particular aspects of a text. (...) The result of an instrumental translation is a text that may achieve the same range of functions as an original text. If the target-text function is the same as that of the source text we can speak of an equifunctional translation; if there is a difference between source and target text functions we should have a heterofunctional translation; and if the (literary) status of the target text within the target-culture text corpus corresponds to the (literary) status the original has in the source-culture text corpus, we could talk about a homologous translation" (Nord 1997, pp. 47–50). Later, Nord emphasizes the independent and autonomous character of instrumental translations: "In the reception of an instrumental translation, readers are not supposed to know they are reading a translation at all" (p. 52).

#### 4.1. Literary Translation Briefs and “Sickert at St Peter’s”

Nord points out that “in the professional practice of intercultural communication, translators rarely start working of their own accord” (Nord 1997, p. 20), but that is precisely what will happen in the present case study. As both the initiator of the communication process and the actual translator, I find myself in the rather unusual position of being able to define all specifications related to the translation of “Sickert at St Peter’s”, of which the present translation-oriented text analysis constitutes a pre-translational part.

Even though the functionalist approach allows plenty of room for all kinds of translation strategies and approaches, I would here like to address what I shall call the literary translation of literary works. This non-pleonastic phrase refers to translations of literature which can be read as if they were also themselves independent works of literature. The objective is to establish a shorthand to easily differentiate between these particular translations from other non-literary translation processes which literary works may also undergo.

The first binary decision to be taken in Nord’s top-down approach is related to the documentary or instrumental function of the translation to be carried out (see note 18). A “document” of a communication carried out by a foreign initiator, sender or author engaged with foreign addressees in the context of a foreign culture seems hardly ideal to engage readers in the target-culture with the artistic, stylistic and not least emotional aspects inherent to a piece of literary writing considered as a work of art. Instrumental translations, as autonomous and independent texts which read as originals and address target-culture readers directly on all levels, are evidently more suitable to this particular task: in addition to producing texts which specifically address the actual intended readership, an instrumental approach also promotes a closing of the subjective distance between the translation and the addressees.

Next comes the desired text function. Downplaying or disregarding the importance of the expressive function in the literary translation of literary works would be completely out of the question, since the use of connotative, expressive and aestheticizing language is a generally accepted major characteristic of literature. As a result, the desired instrumental translation becomes more particularly an equifunctional translation, that is: a translation which strives to keep the function of the original text unaltered. In the literary translation of literary works, this means that the prominent role of the expressive function – which marks the text as a literary – should be maintained in the translation.

Finally we reach the textual level of the translation to be carried out. Here, a corpus-based lexical analysis of the original may provide valuable insights, particularly with regard to style. Given the unwavering importance of style in the reception of literary works, a reenacting of the individual author's style should always be pursued to the highest degree possible in the literary translation of literary works. This particular *Skopos* was explicitly anticipated by Nord:

In der Regel ergibt sich aus dem Befund für die einzelnen lexikalischen Einheiten eines Textes ein "Stilzug" für den gesamten Text. Wenn durch die die Translatfunktion die Wahrung solcher Stilzüge als Übersetzungsziel definiert ist, muss die Übersetzung danach ausgerichtet werden, wie in der ZS der betreffende Stilzug herzustellen ist. Die Analyse der einzelnen Einheiten ist demnach in einen globalen Zusammenhang einzuordnen, in den auch etwa die Befunde aus der Analyse von Inhalt, Aufbau, Syntax etc. integriert werden müssen. (Nord 2009, p. 127)<sup>19</sup>

Seen as a three-step decision process with a strong emphasis on the author's style, the scheme outlined above resembles Katharine Reiß's considerations on the "decisive battle" waged by conscious translators:

Now the *text individual* is placed in the foreground. This analysis is of supreme importance, because the translator's "decisive battle" is fought on the level of the text individual, where strategy and tactics are directed by type and variety. (Reiß 2004, p. 166)

The addressees of the intended translation are here conceived as casual readers, students or scholars of literature who are sufficiently interested in the subject to either know Denton Welch (at least by name) or to welcome the reading of a literary piece from a previously unknown author. The ideal medium/channel to reach these addressees is here envisaged as either a literary magazine or an academic literary journal, where the literary translation of a literary work would most likely be noticed and appreciated by a specialized readership:

---

19 "The analysis of various lexical items in a text can often show that a particular stylistic feature is characteristic of the whole text. If the translation *skopos* requires the preservation of such features, individual translation decisions (in the fields of lexis as well as content, composition, sentence structure, etc.) have to be subordinated to this purpose." (Nord 2005, p. 125)

Literary texts are primarily addressed to receivers who have specific expectations conditioned by their literary experience, as well as a certain command of the literary codes. (Nord 1997, p. 80)

Once the addressees, the medium of communication, the type and the Skopos of the intended translation have all been established, the translation brief is ready and we can proceed to the aspects regarding the results of the corpus analysis and the impact it may have on the planned translation.

#### **4.2. Corpus-based Translation Strategies for “Sickert at St Peter’s”**

Based on the previous discussion, the qualitative result of the corpus-based keyword analysis of “Sickert at St Peter’s” could be summarized as follows:

- The narrator has a keen interest in physical objects;
- Sickert is portrayed as an eccentric character;
- The idea of a shared artistic identity between Sickert and the narrator is suggested;
- Sickert reacts with a grunt each time Raven makes his tacit request for a free oil painting of his mother;
- The word “pictures” appears prominently in the text and contributes to the suggestion of a pervading artistic environment;
- The dialogue is written in a slightly informal style;
- There is a relative lack of subtlety underlying the narrative.

These observations could be implemented as concrete guidelines in order to produce a literary translation of the narrative if the translator accordingly opts to:

- avoid using hyperonyms in the description of objects;
- make use of vocabulary which does justice to Sickert’s eccentricity;
- use first-person plural pronouns whenever possible, so as to suggest the artistic proximity between Sickert and the narrator;
- reproduce the replay effect of Sickert’s grunt by consistently employing a single word to translate “grunt” on all occasions;
- possibly translate the word “pictures” for “painting”, since this would be perfectly in keeping with the broader artistic perspective in the original text;
- use contractions or other slightly informal devices in the translation of dialogues;
- avoid using modals, since the original is quite straightforward.

These suggestions are not to be read as exhaustive, but in my view offer a clear illustration of how a corpus-based approach to Nord's translation-oriented text analysis could be beneficial and reveal textual features which could otherwise go unnoticed – as well as a concrete working method, should the translator wish to implement a corpus approach in his or her professional practice.

## 5. Concluding remarks

With this article, I hope to have demonstrated some of the benefits that a corpus approach may bring to the literary translation of literary works and the theoretical reasons why such an approach is not only legitimate, but also useful for providing a more objective understanding of the inner workings of literary texts. As stated above, computer-processed corpus metrics should never be taken at face value, but contextualized interpretations of corpus analyses can and do provide the translator with powerful insights regarding the text to be translated. Since this approach can be integrated into Nord's translation-oriented text analysis without any sort of theoretical conflict, a computer-assisted step in the pre-translational stage leading up to a literary translation becomes a valuable tool capable of revealing critical stylistic features of the text to be translated. The way in which this sort of analysis incorporates actual textual items as part of a text's "information" (as the concept is understood by Vermeer and Reiß, to whom information is not only related to textual meaning, but also to textual form and effect<sup>20</sup>) seems particularly relevant for a literary – and therefore stylistic – approach to literary translation.

## References

- Anthony, L. (2014). *AntConc* (Version 3.4.4w) [Computer Software]. Tokyo: Waseda University. Available from: [www.laurenceanthony.net/software](http://www.laurenceanthony.net/software). Retrieved on: March 30, 2018.
- Culpeper, J. (2002). Computers, Language and Characterization: An Analysis of Six Characters in *Romeo and Juliet*. In U. Melander-Marttala, C. Ostman & Merja Kyto (Eds.), *Conversations in Life and in Literature: Papers from the ASLA Symposium*

---

20 See Vermeer & Reiß (1991), pp. 58–67, particularly p. 61 and 66.



- (pp. 11–30). Association Suédoise de Linguistique Appliquée (ASLA). Uppsala: Universitetstryckeriet.
- De-la-Noy, M. (1984). Introduction. In M. De-la-Noy (Ed.), *The Journals of Denton Welch*. Penguin.
- De-la-Noy, M. (1987). Introduction. In M. De-la-Noy. *Fragments of a Life Story: The Collected Short Writings of Denton Welch*. Penguin.
- Hermann, J. B., Van Dalen-Oskam, K. & Schöch, C. (2015). Revisiting Style, a Key Concept in Literature. *Journal of Literary Theory (JLT)*, 9 (1), 25–52.
- Mahlberg, M. (2010). Corpus Linguistics and the Study of Nineteenth-Century Fiction. *Journal of Victorian Culture*, 15 (2), 292–298.
- Maiwald, P. (2011). Exploring a Corpus of George MacDonald's Fiction. *North Wind: A Journal of George MacDonald Studies*, 30, Article 5.
- McIntyre, D. & Walker, B. (2010). How Can Corpora Be Used to Explore the Language of Poetry and Drama? In A. O'Keefe & M. McCarthy (Eds.). *The Routledge Handbook of Corpus Linguistics* (pp. 516–530). London: Routledge.
- Nord, C. (2014). *Hürden-Sprünge. Ein Plädoyer für mehr Mut beim Übersetzen*. BDÜ.
- Nord, C. (2009). *Textanalyse und Übersetzen*. Tübingen: Julius Groos (4<sup>th</sup> actual. ed.).
- Nord, C. (2005). *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis*. Translated from the German by Christiane Nord and Penelope Sparrow. Rodopi (2<sup>nd</sup> ed.).
- Nord, C. (1997). *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.
- Reiss, K. (2000). Type, Kind and Individuality of Text: Decision Making in Translation. In L. Venuti (Ed.), *The Translation Studies Reader* (pp. 160–171) S. Kitron (Trad.). London & New York: Routledge.
- Someya, Y. (1998). *e\_lemma\_no\_hyphen.txt*. Available from: <[www.laurenceanthony.net/software/antconc](http://www.laurenceanthony.net/software/antconc)> Retrieved on: March 30, 2018.
- Vermeer, H. & Reiß, K. (1991). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: Max Niemeyer (2<sup>nd</sup> ed.).
- Welch, Denton (1942). Sickert at St Peter's. *Horizon*.

[submitted on March 30, 2018 and accepted for publication on November 12, 2018]

## **EM VIDA E NA HORA DA MORTE TAMBÉM: O QUE DIZEM REGISTROS DE ÓBITO OITOCENTISTAS DA FREGUESIA DE NOSSA SENHORA DA PENHA DE CORUMBÁ (1847–1855)**

IN LIFE AND ALSO AT THE TIME OF DEATH: WHAT DO NINETEENTH CENTURY DEATH RECORDS FROM THE PARISH OF NOSSA SENHORA DA PENHA DE CORUMBÁ TELL (1847–1855)

Maria Helena de Paula\*  
mhp.ufgcatalao@gmail.com

Amanda Moreira de Amorim\*  
amandamoreiradeamorim@gmail.com

A história do Brasil está profundamente atrelada a um regime escravocrata, responsável pela importação de mão de obra africana escrava para o país, que vigorou por aproximadamente quatro séculos. Como ocorreu em grande parte do Brasil, o estado de Goiás teve destacado papel na história da escravidão africana neste país, a comprovarem diversos manuscritos de naturezas variadas, que mencionam escravos que viveram no local. Este artigo propõe analisar histórica e linguisticamente um livro de registro de óbitos, composto por documentos eclesiásticos exarados entre 1847 e 1855, sobre a Freguesia de Nossa Senhora da Penha de Corumbá, visando a estabelecer relações entre a expectativa de vida dos escravos, dos libertos, bem como dos livres, com base em dados como gênero e idade dos falecidos. Para realizar o proposto, elaboramos um inventário com informações referentes aos sujeitos descritos nos registros, baseando-nos no modelo proposto por Santos e Paula (2014) e cotejamos os dados obtidos com estudos que abordam o período escravocrata brasileiro, como os de Libby e Paiva (2005), Paiva (2014), Salles (1992) e outros. Assentado em uma perspectiva de interface dos estudos de Filologia, Linguística e História, este estudo busca compreender aspectos históricos e culturais registrados nos documentos e que sobremaneira os caracterizam.

**Palavras-chave:** Escravidão em Goiás. Filologia. História. Óbitos.

---

\* Universidade Federal de Goiás – UFG, Catalão-Goiás, Brasil. O presente estudo foi desenvolvido sob os auspícios da Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), na forma de auxílio de infraestrutura a jovens pesquisadores e de bolsa de estudos de pós-graduação, respectivamente.

The history of Brazil is closely linked to a slavery regime, responsible for importing African slave labor into the country, which was in force for approximately four centuries. As occurred in many parts of Brazil, the state of Goiás also played an important role in the history of African slavery, proved by various manuscripts of different kinds, which mention slaves who lived in the region. This article aims to analyze both historically and linguistically a book of death records consisting of ecclesiastical documents, written down from 1847 to 1855, concerning the Parish of Nossa Senhora da Penha de Corumbá. The aim is to establish relations between the life expectancy of black slaves, those released and those freed, based on data such as gender and age of the deceased. In order to achieve the objective, we compiled an inventory with information about the people described in the records, based on the model proposed by Santos and Paula (2014) and compared the data obtained with studies that discuss the slavery period in Brazil, such as Libby and Paiva (2005), Paiva (2014), Salles (1992), among others. Based on an interface perspective of the studies of Philology, Linguistics and History, this study attempts to understand historical and cultural aspects recorded in the documents, and what characterizes them.

**Keywords:** Slavery in Goiás. Philology. History. Death.



## 1. Introdução

Durante 388 anos, vigorou no Brasil um regime escravocrata, responsável pela importação de mão de obra africana escrava para o país. O atual estado de Goiás teve destacado papel na história da escravidão brasileira, visto a abundância de documentos manuscritos encontrados na região, de natureza cartorial, judiciária e eclesiástica, os quais fazem menção aos escravos que viveram no local.

Parte desses manuscritos se encontra nos arquivos digitais dos projetos “Formação de *corpora* escritos de Goiás – leitura e edição de documentos” e “Em busca da memória perdida: estudos sobre a escravidão em Goiás”, que vêm sendo organizados desde 2007 e constam no acervo digital do Laboratório de Estudos do Léxico, Filologia e Sociolinguística (LALEFIL), da Universidade Federal de Goiás, na cidade de Catalão, Brasil. A presente contribuição se insere no campo das Humanidades, em específico na interface dos estudos de Filologia, Linguística e História e o acervo digital de que este estudo se vale vem sendo constituído a partir de técnicas de

digitalização, que prescindem de aparelhos eletrônicos que emitem calor ou luz, com o fito de preservar o *status* material do documento.

Para a realização dessa pesquisa, selecionamos como objeto de estudo um livro de registro de óbitos, extraído dos arquivos supramencionados, o qual é composto por documentos eclesiásticos exarados entre os anos de 1847 e 1855, na Freguesia de Nossa Senhora da Penha de Corumbá, na atual cidade de Corumbá de Goiás, redigidos pelo Vigário Manoel Innocencio da Costa Campos. Tais documentos apresentam informações características de cada um dos falecidos, como: seus nomes completos, a idade em que foram a óbito, seu estado conjugal, entre outras. Com base nessas informações, intentamos estabelecer relações entre a expectativa de vida dos escravos, dos libertos, bem como dos livres, levando em consideração o gênero dos falecidos e outras informações, por meio da realização de um inventário do livro de registros.

Para tanto, iniciamos nosso percurso com a leitura minuciosa do livro indicado, visando à seleção dos dados no *corpus* de estudo. Em seguida, elaboramos um inventário composto de informações referentes aos sujeitos descritos nos registros, disposto em forma de tabela. Após o levantamento das informações, comparamos os resultados obtidos com obras que versam acerca do período escravocrata brasileiro, de estudiosos como Libby e Paiva (2005), Paiva (2014) e Salles (1992), que disserta especificamente sobre a escravidão em Goiás. Pretendemos, assim, demonstrar, por meio de gráficos e tabelas, as taxas de mortalidade na Freguesia de Nossa Senhora da Penha de Corumbá, em um recorte de aproximadamente 08 anos (entre 1847 e início de 1855), embasando-nos, sobretudo, em um estudo linguístico realizado a partir de lexias extraídas do livro em pauta.

A relevância desta pesquisa pauta-se no acesso, conhecimento e publicação de aspectos linguísticos, históricos e culturais da sociedade brasileira oitocentista, posto que os manuscritos analisados revelam traços históricos e linguísticos importantes da época em que foram exarados, o que proporciona o entendimento de parte da história da escravidão negra em Goiás, além de viabilizar futuros estudos acerca da história do Brasil e sua configuração na escravidão africana no mundo.

Pleiteando apresentar os resultados obtidos na pesquisa em tela, este artigo se dividirá em três seções. A primeira contextualiza, brevemente, a instauração da escravidão no Brasil e no estado de Goiás, enquanto o segundo segmento trata dos caminhos teóricos que nortearão a análise dos dados, que será desenvolvida na terceira seção.

## 2. Uma breve história da escravidão no Brasil

A instauração de uma colônia portuguesa em terras brasileiras trouxe consigo um sistema de trabalho compulsório, como apontam Libby e Paiva (2005), o qual levou inúmeros indígenas, africanos e mestiços a serem escravizados, posto que a prática escravagista era o principal meio de se gerar mão de obra estável e barata e atender ao objetivo central dos colonizadores, pautado no crescimento econômico local.

Contudo, é importante ressaltar que tais práticas existem desde o período pré-histórico e relatos podem ser observados em diversas sociedades. Malheiro ([1866] 2014, p.13), em sua obra “A escravidão no Brasil”, pontua que “a escravidão antiga achava sua escusa no direito do vencedor em guerras internacionais”. Desta maneira, povos que eram vencidos em guerras contra seus rivais, além de aprisionados, tornavam-se escravos, atitude empregue para se poupar a vida dos vencidos em batalha.

Um exemplo bastante conhecido de práticas escravagistas antigas ocorria no Império Romano, civilização que perdurou entre 27 a.C. e 476 d.C. Nessa sociedade, havia diferentes modos de se escravizar uma pessoa, todos eles legitimamente reconhecidos, como o roubo em flagrante, em que “oladrão (*furmanifestus*) era açoitado e entregue como escravo ao ofendido” (Malheiro, [1866] 2014, p. 13). Nesses casos, os escravos eram, predominantemente, brancos e sua mão de obra era empregada na agricultura comercial, nos transportes marítimos, na mineração e nos ofícios artesanais, conforme expressam Libby e Paiva (2005).

Esse sistema sofreu alterações até sua chegada no Brasil, no século XVI. Em um primeiro momento, os colonos portugueses priorizaram a escravização dos povos nativos, que já habitavam a região por eles colonizada. Todavia, a escravidão indígena não se mostrou satisfatória, conforme expressa Xavier (2010) e, em 1702, a Carta Régia ao Governador do Maranhão proibia o cativeiro dos índios, permitindo, entretanto, sua administração por tempo limitado. A autora aponta que essa administração deveria ocorrer de maneira remunerada, condição não respeitada por muitos. Isso, então, desencadeou a exploração indígena, uma escravização disfarçada, que provocou inúmeras mortes de nativos, decorrentes de jornadas de trabalho intensas e doenças trazidas de outras regiões.

A resistência dos senhores para a aquisição de escravos africanos se deu, inicialmente, motivada pelos altos valores por que eram vendidos, em decorrência de sua importação. Salles (1992), em seu livro “Economia e escravidão na Capitania de Goiás”, resalta que, no começo do século XVII, um escravo africano poderia custar o equivalente a quatro cativos indígenas, o

que ratifica a predileção senhorial pelos escravos nativos. Entretanto, alguns aspectos como a força física do negro e, principalmente, os futuros lucros oriundos de sua aquisição, tornaram-se grande atrativo e, progressivamente, constatou-se um maior número de escravos africanos em terras brasileiras.

Na Capitania de Goiás, a exploração aurífera foi a principal atividade que movimentou a ocupação e o povoamento da região, segundo Salles (1992). Documentos apontam que em 1752 o primeiro comboio de negros foi registrado em Goiás sem, contudo, informações acerca do montante de escravos. A autora ressalta, ainda, que a descoberta de algumas minas, como as de Jaraguá, Tesouras e Cocais, decorre de explorações de escravos africanos que trabalhavam neste meio. Por se tratar de atividades de mineração e exploração aurífera, os escravos “minas” tornaram-se os preferidos nessa região, posto que possuíam experiência na mineração das costas africanas. Mas, além disso, a mão de obra escrava também foi empregada na agricultura e na pecuária.

Os rastros deixados pelos mancípios que atuaram no território goiano encontram-se documentados em diversos manuscritos da época em questão, exarados em diferentes cidades goianas, como Catalão, Jataí, Luziânia (antiga Santa Luzia), Silvânia (antiga Bonfim), e de variadas tipologias, como os registros de batizado e os de óbito, as escrituras públicas de compra, venda, doação, hipoteca ou troca de escravo, os registros, cartas ou escrituras de liberdade, entre outros.

Neste estudo que ora apresentamos, os dados e sua análise são de um livro de registro de óbitos, exarado na freguesia de Corumbá de Goiás, região povoada em decorrência do descobrimento de minas. Nosso recorte limita-se aos anos de 1847 e 1855, quando a atividade mineradora encontrava-se extenuada; no entanto, a utilização de trabalho escravo mantinha-se firme na região, com sua mão-de-obra podendo ser aproveitada em atividades como agricultura, pecuária, serviço doméstico, artesanato e outros.

### 3. Na trilha da história: Os caminhos teóricos

Este estudo tem como *corpus* 658 registros de óbitos, documentos manuscritos que datam do século XIX, provenientes da Freguesia de Nossa Senhora da Penha de Corumbá, exarados pelo Vigário Manoel Innocencio da Costa Campos. Tais documentos são registros paroquiais, constituintes de acervos encontrados em igrejas católicas, cuja tipologia é definida por Bellotto (2002, p. 84) como “documento diplomático testemunhal de assentamento”. A autora versa, em sua obra, acerca da diferença entre tipo e espécie documental. A

espécie documental se refere à disposição e à natureza das informações contidas no documento, obedecendo a fórmulas previamente convencionadas, enquanto o tipo documental trata-se da “configuração que assume a espécie documental de acordo com a atividade que a gerou (...)” (Bellotto 2002, p. 19).

Desta forma, os registros de óbito analisados nesse estudo correspondem à espécie documental diplomática e, quanto ao tipo, situam-se no campo dos documentos eclesiásticos, posto que obedecem a normas pré-estabelecidas pela Igreja Católica para suas composições. Atendem ao subtipo registro, já definido como documento diplomático testemunhal de assentamento, e ao subtipo óbito, uma vez que é possível localizar diferentes registros em documentos eclesiásticos, como os registros de batismo e casamento.

Ao realizarmos leitura acurada dos manuscritos selecionados para a composição do *corpus*, deparamo-nos com padrões que se repetem a cada um deles, a saber: abertura com a data de registro; o nome do declarante do óbito; seu local de moradia; data, local, horário e causa do óbito; nome do falecido; suas características, como cor da pele, condição social (escravo/forro); idade em que faleceu; estado conjugal, acompanhado do nome do cônjuge, quando casado; nome dos pais; ritual recebido; local de sepultamento e, por fim, nome do vigário responsável pelo registro. Por vezes, algumas destas informações eram ignoradas pelo declarante, como visto repetidamente na causa do óbito e no nome dos pais. Por outro lado, alguns registros trazem informações complementares, como o ofício exercido pelo falecido e, em caso de escravo ou forro, o nome de seu senhor.

A partir dessas informações, elaboramos um inventário, baseando-nos no modelo proposto por Santos e Paula (2014), no trabalho “Escravidão em Goiás: mortalidade branca e escrava na Vila de Santa Luzia entre os anos de 1786 a 1814”. Ao adaptarmos o modelo supradito, nosso inventário foi composto pelos seguintes dados: número do documento; fôlio em que se localiza no códice; data de registro; data do óbito; nome do falecido; características; dono (em caso de escravo); idade; estado conjugal; sacramentos; profissão; nome do pai; nome da mãe; motivo do óbito; local do enterro; vigário responsável pelo registro. Por ser o *corpus* composto por 658 registros, este extenso inventário foi base para as ilustrações ao longo do estudo, não constando, portanto, no corpo deste texto.

Optamos pela categoria *características*, do mesmo modo que Santos e Paula (2014), porque esta abrange um maior número de informações acerca do falecido. Neste campo, arrolamos dados como *qualidade*, *condição social* e *origem*. Para conceituar *qualidade*, apoiamo-nos na perspectiva de Paiva (2014), em sua tese “Dar nome ao novo: uma história lexical da

Ibero-América, entre os séculos XVI e XVIII (as dinâmicas de mestiçagens e o mundo do trabalho)”. O autor emprega a referida lexia para classificar os sujeitos, diferenciá-los e hierarquizá-los, através de sua origem familiar, seus traços fenotípicos e suas condições sociais. Entretanto, a atribuição das qualidades não ocorria de maneira uniforme, posto que dependia, principalmente, “do olhar individual de cada pessoa e das conveniências, o que permitia que uma pessoa pudesse ter suas qualificações alteradas ao longo dos anos” (Almeida *et al.* 2017, p. 162).

Como *condição social*, destacamos se o falecido era *escravo*, *liberto* (também apontado como *forro*) ou *livre*. Neste ponto, faz-se mister ressaltar a distinção entre as lexias *liberto* e *livre*. *Liberto*, no contexto escravagista aqui apresentado, refere-se à “(...) nova condição à qual o escravo se submetia ao alcançar a tão almejada alforria” (De Paula & Amorim 2016, p. 138). Ou seja, ao ser alforriado, o escravo tornava-se liberto da situação opressora em que vivia, adquirindo, pois, a liberdade. Entretanto, a lexia *livre* não expressa uma condição adquirida após determinado fato, mas “(...) uma condição de nascença, daquele que já nascia em liberdade” (De Paula & Amorim 2016, p. 139). Deste modo, *liberto* (ou *forro*) refere-se àquele que foi escravo de outrem, conforme observa-se nos registros destes sujeitos, os quais, na maioria das vezes, fazem menção ao seu antigo senhor, comprovando sua origem social. Por outro lado, *livre* refere-se à condição nata do sujeito que nascia fora do berço da escravidão.

Acerca da *origem*, listamos no inventário as lexias que se referem à nação do sujeito registrado. Contudo, as lexias utilizadas nos documentos podem não demonstrar, de maneira correta, o local de origem do indivíduo, sobretudo em se tratando dos escravos. Barros (2014) apresenta em seu livro “A construção social da cor: diferença e desigualdades na formação da sociedade brasileira” as noções de etnia de origem e etnia do tráfico. A primeira caracteriza a etnia primeira dos escravos, sinalizando seu local de nascimento, como poderia ocorrer com os escravos identificados como *angola*, em alusão ao país localizado na costa ocidental africana de onde teria sido trazido. Já a segunda provém de uma atribuição dada aos escravos por seus comerciantes, com o principal objetivo de demarcar os tipos de serviços para os quais eram mais bem indicados, como poderia acontecer com um escravo denominado *mina*, o que demarcaria sua adequação para trabalhar com exploração de minérios, e não necessariamente sua origem na Costa da Mina, região localizada no Golfo da Guiné.

O período em que foram exarados os documentos que compõem nosso *corpus* de estudo, entre 1847 e 1855, permeia a promulgação da Lei Eusébio de



Queirós, de 1850, que colocava fim ao tráfico internacional de escravos, proibindo o comércio entre os continentes americano e africano. Desse modo, ainda que escravos estrangeiros chegassem ao Brasil, sua etnia de origem era, por vezes, suprimida e, em decorrência disso, encontramos em nosso inventário apenas dois escravos registrados como *africanos*, um registrado como *nagô* e um como *angola*, todos com idade superior a 60 anos.

Para a análise dos dados retromencionados, dadas as interfaces disciplinares que caracterizam este estudo, fundamentamo-nos, principalmente, na teoria proposta por Spina (1977). Para o autor, a Filologia cumpre três funções, a saber: *função substantiva*, que visa à restituição do texto a sua forma genuína, para sua publicação; *função adjetiva*, a qual deduz do texto aquilo que não está evidente nele, como sua datação, por exemplo; e a *função transcendente*, que busca depreender as relações históricas que motivaram o texto.

Com o intuito de trazer a lume importantes aspectos da história e da cultura do Brasil oitocentista, servimo-nos da função transcendente, posto que “(...) o texto deixa de ser um fim em si mesmo da tarefa filológica, para se transformar num instrumento que permite ao filólogo reconstituir a vida espiritual de um povo ou de uma comunidade em determinada época” (Spina 1977, p. 77). Desta maneira, buscamos, através do texto (aqui, o conjunto dos registros de óbito estudado), compreender as relações que o motivaram historicamente, além de perscrutar a história e a cultura da época.

Para além da teoria filológica, apoiamo-nos, também, na teoria lexical, uma vez que “(...) o léxico está prenhe de informações históricas das civilizações presentes em textos orais e/ou escritos, haja vista que ele é o responsável pela representação do real na língua, intermediando, assim, a relação do homem com o seu meio” (Xavier 2012, p. 470). Podemos, então, depreender que o léxico de uma língua carrega importantes traços sociais e culturais – em nosso caso, do período escravocrata na província de Goiás, o que pode ser percebido a partir das lexias que tratam sobre escravos, senhores e suas relações. Noutras palavras, porque “o léxico é o repositório mais dinâmico das configurações culturais denotadas em uma dada língua” (De Paula 2007, p. 49) é nele que podemos encontrar os indícios dos feitos históricos e práticas culturais da época nestes documentos.

Portanto, é essencial para a construção de nossa análise o suporte teórico de obras voltadas para a História, posto que “língua, história e cultura caminham sempre de mãos dadas e para conhecermos cada um desses aspectos, faz-se necessário mergulhar nos outros, pois nenhum deles caminha sozinho e independente” (Abbade 2008, p. 716).

#### 4. Os registros de óbito da freguesia de Nossa Senhora da Penha de Corumbá

A Igreja Matriz de Nossa Senhora da Penha de França localiza-se, hoje, na cidade de Corumbá de Goiás, e desempenhou importante papel de entreposto entre as capitanias de Goiás, Minas Gerais e Bahia. Em 2004, a Igreja foi tombada como Patrimônio Cultural Brasileiro pelo *Instituto do Patrimônio Histórico e Artístico Nacional* (IPHAN), assim como o Conjunto Arquitetônico de Corumbá de Goiás. Sua construção iniciou-se ainda na primeira metade do século XVIII e remete à arquitetura vernacular, em decorrência do sistema construtivo empregado nos primórdios da colonização do Centro-Oeste brasileiro.

Nos tópicos seguintes, tecemos algumas observações acerca dos dados levantados no inventário do livro de registros de óbitos da referida Igreja Matriz, sob os quais se fundamentam nossas análises.

##### 4.1. Número de óbitos por ano

O livro de registros de óbitos selecionado para este estudo compõe-se de 658 documentos, exarados entre janeiro de 1847 e janeiro de 1855, o que nos proporciona um recorte de, aproximadamente, oito anos. Após a realização do inventário do livro, de que extraímos dados relativos a cada registro de óbito, chegamos aos seguintes números de mortes por ano:

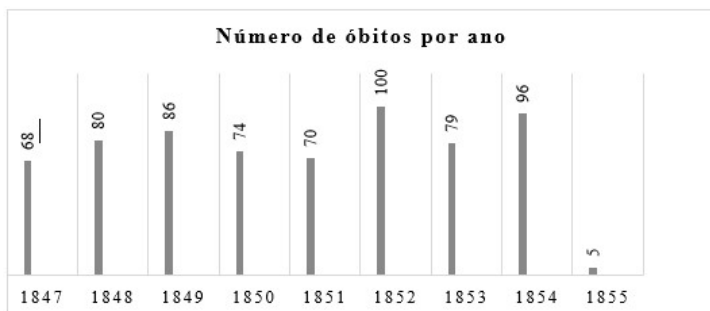


Figura 1. Números de óbitos por ano.

Fonte: elaborado pelas autoras.

Com base no total de documentos arrolados no inventário, estabelecemos uma média de 82 mortes registradas por ano. Entretanto, o ano de

1852 mostrou-se o mais expressivo, conforme o gráfico acima, ao totalizar 100 óbitos. Em contrapartida, em 1855 registraram-se apenas 05 mortes, em decorrência de seus dados limitarem-se ao dia 16 de janeiro.

#### 4.2. Número de óbitos de escravos, libertos e livres

Em relação à estrutura social vigente no período escravocrata oitocentista, especialmente na província de Goiás, local em que os registros de óbito foram exarados, diferenciamos o número de óbito dos livres, dos escravizados e dos libertos, conforme diferenciação entre *livre* e *liberto*, apresentada no tópico 3. A divisão do número de mortes de livres, escravizados e libertos nos permitiu elaborar o seguinte gráfico:



Figura 2. Gráfico ilustrativo do número de óbitos de escravos, libertos e livres.

Fonte: elaborado pelas autoras.

Esses dados nos revelam o baixo percentual de mancipios em Corumbá de Goiás, apenas 0,87% no período estudado. Salles (1992) expressa que a descoberta do ouro em minas goianas foi fator determinante para o desenvolvimento social e econômico da região. A demanda de mão de obra para o trabalho no garimpo foi a principal responsável pela importação de escravos africanos para a região, razão pela qual a ocupação negra em Goiás se deu de forma regular e constante. Contudo, a exploração descontrolada enfraqueceu a atividade, gerando a decadência das minas e, no fim do século XVIII e início do século XIX, alguns escravos foram remanejados para as zonas açucareiras paulistas e fluminenses.

Os escravos que permaneceram na região foram remanejados para novas atividades, podendo ter sido empregados na formação das primeiras lavouras, bem como os serviços domésticos, no artesanato, nas fábricas de

açúcar e também no transporte, conforme indica Salles (1992). Em nosso inventário, encontramos 02 registros de escravos da casa, o que indicava a realização de serviços domésticos, sendo os únicos registros em que o vigário responsável demarcou a ocupação do escravo.

#### 4.3. Número de óbitos de acordo com as características

Na categoria característica, obtivemos o seguinte resultado, em relação ao número de óbitos:

Tabela 1. Número de óbitos de acordo com as características.

	Livres	Escravos	Forros	Não especificado	Total
Branços	187	-	-	-	187
Pardos	351	12	02	-	365
Crioulos	22	51	06	-	79
Cabras	01	07	-	-	08
Angola	01	-	-	-	01
Caboclo	01	-	-	-	01
Mestiça	-	01	-	-	01
Nagô	-	01	-	-	01
Não especificado	08	03	-	04	15

Fonte: elaborada pelas autoras.

As lexias acima descritas revelam a diversidade do Brasil escravocrata, especialmente na região do estado de Goiás, considerando que em 1850 a lei Eusébio de Queirós proibiu o tráfico de escravos para o Brasil. A unidade lexical *branco* está diretamente relacionada à condição jurídica elevada do sujeito registrado, no caso *livres*, posto que uma grande parcela deles descendia dos colonizadores portugueses. Por vezes constatamos, no decorrer do livro de óbito, diferenças entre os registros dos livres em relação aos demais, sendo mais completos, com informações adicionais.

Já *pardo* era uma categorização constantemente utilizada, com ampla gama de significados. Em registros datados do século XVI, seu uso indicava a miscigenação entre povos distintos, a saber: negros, crioulos, mulatos ou zambos com brancos ou índios. Concomitantemente, representava a cor de pele situada entre o branco e o preto, fruto da mescla entre os povos retrorreferidos. No século

XVIII, *pardo* ganha uma nova acepção, passando a representar os nascidos de escravos forros. Almeida *et al.* (2017, p. 167) afirmam que a *lexia* era empregada “como um indicador social, provocando uma hierarquização interna entre os forros e seus descendentes, pois como cor da pele, o pardo aproximava-se mais do branco, modelo ideal a ser seguido na época em questão”.

Nos documentos aqui analisados, há poucos registros de pardos escravos. Deste modo, inferimos que seu uso era recorrente para denominar os nascidos livres no Brasil, filhos de pais escravos. Acerca da cor da pele, no registro de número 141, deparamo-nos com a designação “pardo moreno”, o que indica que, além de nascido no país, o falecido possuía a pele de cor correspondente àquela entre o branco e o preto. O “Diccionario da Lingua Brasileira”, de Luiz Maria da Silva Pinto (1832), define moreno como “de cor parda escura”, o que corrobora o exposto.

Assim como *pardo*, a *lexia crioulo* também refere-se ao nascido no Brasil. Raphael Bluteau (1712-1728, v. II, p. 613), em seu “Vocabulario Portuguez e Latino”, conceitua crioulo como o escravo nascido na casa de seu senhor. Contudo, identificamos em nosso *corpus* 02 registros de sujeitos caracterizados como “crioulo africano”, ambos antigos escravos, já libertos. Essa combinação de qualidades distintas expressa uma alteração no sentido primeiro de *africano*, que se refere aos escravos oriundos da África para servirem como mão de obra no Brasil.

Entretanto, presumimos que a junção de *crioulo* à qualidade retrorreferida não expressa que estes alforriados nasceram em países da América do Sul, e a idade avançada de 90 anos de um dos registrados corrobora esta assertiva. Almeida (2017) expressa que as categorizações encontradas em diversos documentos públicos, que dissertam sobre os escravos, são responsáveis, na maioria das vezes, por decretar as condições de vida a que esses sujeitos estariam expostos. De acordo com a autora, “isso se justifica porque a cada uma dessas qualificações era conferido um lugar na sociedade, estando os africanos na base dessa pirâmide hierárquica enquanto o pardo encontrava-se numa linha de transição entre a escravidão e a liberdade” (Almeida 2017, p. 523). Consideramos, então, que os *africanos* registrados em nosso *corpus* adquiriram a qualidade de *crioulo* ao serem contemplados com a liberdade, tendo em vista que ser apenas africano tornava-os “inferior em relação aos escravos nascidos no Brasil” (Almeida 2017, p. 487), ou indicar o caminho reverso, sendo o *crioulo* caracterizado como *africano* quando submetido ao ato da escravidão. Situação semelhante ocorre com uma escrava registrada como *preta forra parda*, em que a alforriada recebia a alcunha de *parda*, em complemento a *preta* (utilizado muitas vezes como sinônimo para *africano*), possivelmente ao receber sua liberdade.

Já a lexia *cabra* sofreu variações de sentido no decorrer dos anos. No século XVI, era empregada para designar, de forma pejorativa, os índios nativos, assim como os filhos mestiços nascidos da mescla entre índios e africanos, índios e negros, mulatos e negros e negros e brancos, conforme indicam Almeida, Amorim e De Paula (2017). Assim como *pardo*, no século XVIII, sua acepção passa a abranger a cor da pele, mas diferentemente daquela, *cabra* designa moreno claro.

*Caboclo* e *mestiço* também indicam miscigenação. A primeira lexia é definida no dicionário Novo Dicionário Eletrônico Aurélio, de Ferreira (2004), como “mestiço de branco com índio; cariboca, carijó”, enquanto a segunda é conceituada por Bluteau (1712-1728, p.455) como pessoa nascida de pais de diferentes nações. Por outro lado, *angola* e *nagô* não apontam misturas étnicas, mas a etnia de origem desses sujeitos, provenientes do território africano.

#### 4.4. Os locais de enterro

No que tange aos locais de enterro registrados no livro de óbito aqui analisado, obtivemos os seguintes resultados, organizados na tabela abaixo:

Tabela 2. Os locais de enterro na Matriz de Nossa Senhora da Penha de Corumbá.

Locais de enterro	Livres	Escravos	Forros	Não especificado	Total
Nesta Matriz	522	65	07	06	600
Dentro desta Matriz	36	03	-	01	40
Não especificado	-	04	-	05	09
Varanda de São Lesbão	02	02	-	-	04
Adro desta matriz	01	01	-	-	02
Capela dos Angicos	01	-	-	-	01
Fora desta Matriz	01	-	-	-	01
Varanda desta Matriz	-	-	01	-	01

Fonte: elaborada pelas autoras.

As Constituições do Arcebispado da Bahia, redigidas por D. Sebastião Monteiro da Vide (1853), livro formulado como parâmetro para a vida religiosa no Brasil Colônia, expressa que era um costume louvável para as Igrejas Católicas realizar o enterro dos cristãos nas próprias igrejas, posto que ali diversos fiéis compareceriam para ouvir e assistir a missas e outras celebrações religiosas, lembrando-se sempre de encomendar a Deus as almas dos defuntos enterrados. Em nosso inventário, localizamos 40 registros de sepultamentos dentro da Matriz, o que atesta este costume.

Contudo, ao separarmos este número com base na condição social do sujeito, constatamos apenas 3 escravos enterrados dentro da Matriz. Vide (1853, p. 295) indica que alguns senhores mandavam “(...) enterrar seus escravos no campo, e matto, como se forão brutos animaes”. Ainda que houvesse multas e penalidades para os que praticavam tal ato, o baixo número de escravos sepultados demonstra que esta conduta não era totalmente inexistente.

Era direito expresso no livro supracitado que todo cristão escolhesse o local de sua sepultura, fosse na Igreja ou no adro, conforme sua vontade e sua devoção. Entretanto, no número mais expressivo de registros, observamos a recorrência da seguinte informação: “foi sepultado nesta Matriz” (fólio 1v), o que não nos informa com exatidão a localização da sepultura, podendo ser dentro, fora ou no adro da Matriz em questão.

De acordo com o “Glossário de termos sobre religiosidade”, de Nunes (2008, p. 19), o adro era definido como “lugar aberto na frente ou ao redor das igrejas, de ordinário resguardado por muros baixos. Antigo cemitério quando os enterramentos eram feitos junto aos templos”. Em nosso inventário, há dois (02) registros de sepultamentos no adro da Matriz; um de escravo e um de livre. Araújo (2011, p. 8) aponta que “aos forros, livres pobres e escravos restavam o adro, parte que circunda a igreja”, o que nos indica que a parda ali enterrada era livre e pobre. O mesmo é indicado nos registros em que os sujeitos foram sepultados fora da Matriz e na varanda, posto que, apesar de fazer parte do solo sagrado, o lado exterior da igreja encontrava-se mais propício a atos de profanação (Araújo 2011). Quanto à Varanda de São Lesbão e à Capela dos Angicos, não obtivemos material teórico suficiente que fundamentasse uma análise com os dados inventariados.

#### 4.5. Sacramentos

Na categoria sacramentos, obtivemos os seguintes dados, de acordo com o número de sujeitos registrados no livro de óbito analisado:

**Tabela 3. Os sacramentos.**

Sacramentos	Livres	Escravos	Forros	Não especificado	Total
Batizado	01	-	-	-	01
Sacramentado	20	01	-	-	21
Não especificado	542	74	08	12	636

**Fonte:** elaborada pelas autoras.

De acordo com o “Glossário de termos sobre religiosidade” (Nunes 2008), há sete sacramentos na religião cristã: o batismo, a confirmação, a eucaristia, a penitência, a unção dos enfermos, a ordem e o matrimônio. À exceção do sacramento da ordem, que conferia o poder e a graça de exercerem funções de cunho eclesiástico, todos os outros deveriam ser realizados em qualquer pessoa, fosse ela livre, forra ou escrava, sob pena de punição para aqueles que não chamassem o padre para o sacramento final, conforme expressa Vide (1853).

Esporadicamente, deparamo-nos com expressões do tipo “recebeu os sacramentos” (fólio 81r), “recebeu os últimos sacramentos” (fólio 64v) e “foi sacramentado” (fólio 2v). Nas expressões citadas, conjecturamos que o sacramento recebido pelo falecido seja o da extrema-unção, ou unção dos enfermos, já no leito de morte, posto que este melhor se encaixaria como último sacramento.

Santos (2014) afirma que alguns padres eram encorajados a aprender idiomas africanos, com vistas a facilitar o momento de extrema-unção de homens e mulheres africanos que não compreendiam a língua portuguesa. No nosso estudo, não obtivemos semelhante resultado sobre a realização dos sacramentos em escravizados, porque essa informação foi pouco recorrente nos registros inventariados. Há apenas um registro de escrava sacramentada, Severina, pertencente ao Reverendo Vigário, provável motivo pelo qual recebeu os sacramentos, o que corresponde a apenas 4,76% do total, enquanto os livres, sejam eles pardos ou brancos, correspondem a 95,24% do total de sacramentados.



#### 4.6. Idades

Após inventariarmos as idades correspondentes às pessoas registradas no livro de óbito da Freguesia de Nossa Senhora da Penha de Corumbá, elaboramos o seguinte quadro, adaptado do modelo publicado por Santos (2014), em que: **H** corresponde a *homens*; **M** corresponde a *mulheres*; **L** refere-se a *livres*; **E** refere-se a *escravos*; **L/F** refere-se a *libertos/forros*; e “?” corresponde a *não especificado*.

**Tabela4. Número de mortes por faixa etária de acordo com o inventário do livro de óbitos.**

Característica	Gênero	Condição social	Recém-nascidos a 1 mês	1 mês a 18 meses	18 meses a 12 anos	12 anos a 18 anos	18 anos a 40 anos	40 anos a 60 anos	Maiores de 60 anos	Não especificado	Total
Branços	H	L	16	12	09	03	18	17	13	03	91
	M	L	18	07	15	08	13	19	11	05	96
Pardos	H	L	34	35	28	04	41	34	07	20	203
		E	-	03	01	-	01	02	-	-	07
	M	L	16	22	13	07	33	29	17	11	148
		E	01	-	01	02	01	-	-	-	05
		L/F	-	-	-	-	-	01	01	-	02
Crioulos	H	L	-	04	01	-	-	01	03	-	09
		E	01	01	03	02	13	01	03	02	26
		L/F	-	-	-	-	-	02	01	02	05
	M	L	01	-	04	-	03	02	03	-	13
		E	04	02	02	03	06	05	03	-	25
		L/F	-	-	-	-	-	01	-	-	01
Cabras	H	L	-	-	-	-	-	-	01	-	01
	M	E	01	01	01	-	02	01	-	01	07
Angola	H	L	-	-	-	-	-	-	01	-	01
Caboclo	H	L	-	-	-	-	01	-	-	-	01
Mestiça	M	E	-	-	-	-	-	-	-	01	01
Nagô	H	E	-	-	-	-	-	01	-	-	01

Característica	Gênero	Condição social	Recém-nascidos a 1 mês	1 mês a 18 meses	18 meses a 12 anos	12 anos a 18 anos	18 anos a 40 anos	40 anos a 60 anos	Maiores de 60 anos	Não especificado	Total
Não especificado	H	L	01	-	-	-	-	01	-	01	03
		E	-	01	-	-	-	-	-	01	02
	M	L	-	-	-	01	02	01	01	-	05
		E	-	01	-	-	-	-	-	-	01
	?		-	-	-	-	-	-	-	04	04
Total			93	89	78	30	134	118	65	51	658

Fonte: elaborada pelas autoras.

Por questões de clareza e para facilitar o entendimento da tabela, excluímos as linhas cuja ocorrência foi de zero entradas, como ocorreu nas categorias “brancos/homens/escravos” e “brancas/mulheres/forras”, uma vez que durante o período escravocrata brasileiro não há registros de escravização de brancos.

Em relação aos escravos, notamos que o grupo com maior número de mortes registrado era o de crioulos, o que corresponde a 7,75% do total de óbitos inventariados, ao somarmos os valores referentes tanto aos homens quanto às mulheres. Apenas em três registros observamos a informação relacionada à *causa mortis*. Uma escrava de 30 anos faleceu repentinamente, um escravo de 20 anos morreu vítima de picada de cobra e um escravo de 20 anos suicidou-se com cinco facadas no estômago, após ter fugido do sítio em que vivia. Ainda sobre os crioulos escravos, o maior número de óbitos se dá no intervalo de 18 a 40 anos: em se tratando de homens, corresponde a 1,98% do total de mortes registradas; no que tange às mulheres, corresponde a 0,91% dos óbitos inventariados.

No tocante aos pardos escravos, tanto homens quanto mulheres, o número de mortes registradas em nosso inventário corresponde a 1,82% do valor total de óbitos, número relativamente baixo, visto que a maioria dos registros corresponde aos pardos livres. Não encontramos, nos registros, informações referentes às causas dos óbitos desses escravos. Ademais, a faixa etária com o maior número de mortes varia para os homens e as mulheres. A primeira faixa localiza-se no intervalo entre 1 a 18 meses, correspondendo a 0,46% do valor total de óbitos, e a segunda localiza-se no intervalo entre 12 a 18 anos, o que representa 0,30% do total de mortes registradas.

Quanto às escravas classificadas como cabras, obtivemos um total de 1,06% das mortes inventariadas. A maior concentração de óbitos situa-se no intervalo de 18 a 40 anos, o que corresponde a 0,30% do valor total de registros. Acerca da *causa mortis*, apenas um registro apresentou essa informação, em que uma escrava de 50 anos faleceu repentinamente.

Em se tratando do escravo nagô e da escrava mestiça, ambos correspondem à mesma porcentagem, de 0,15%, em relação ao valor total de óbitos. Na categoria “não especificado”, há três registros de mortes de escravos, o que representa 0,45% das mortes inventariadas. Contudo, em nenhum dos registros, seja nagô, mestiça ou “não especificado”, identificamos a *causa mortis*, visto que não constava tal informação.

No que se refere aos libertos e forros, há poucos registros sobre eles, contabilizando-se apenas 1,21% do total de 658 mortes inventariadas. O maior número de óbitos corresponde aos crioulos, com idade entre 40 e 60 anos, computando 0,30% do total. Há apenas um registro com a *causa mortis* delimitada, de uma crioula de 60 anos que faleceu repentinamente.

Já os livres representam a maior porcentagem de óbitos inventariados, correspondendo a um total de 86,78% dos registros. Na categoria brancos, a faixa etária com o maior número de óbitos varia para os homens e para as mulheres. A primeira localiza-se no intervalo entre 18 a 40 anos, referindo-se a 2,74%, e a segunda localiza-se no período entre 40 e 60 anos, correspondente a 2,89% do total. Ao realizar a leitura do livro de óbitos em estudo, constatamos que os registros dos livres eram, por vezes, mais completos de informações. Por este motivo, contabilizamos 10 documentos em que a *causa mortis* é especificada, a saber: 03 pessoas faleceram repentinamente; 02 mulheres faleceram de complicações no parto; 02 pessoas faleceram de *hydropezia*, apresentada por Santos (2014) como hidropsia, problema relacionado a inchaço nas pernas; 1 pessoa morreu queimada; e 02 faleceram por engasgo, uma criança engasgada com um biscoito e uma idosa com mal de engasgo, o que inferimos tratar-se de problemas com engasgo recorrente.

Quanto aos pardos livres, a faixa etária com maior número de óbitos localiza-se no intervalo entre 18 e 40 anos, tanto para homens quanto para mulheres, o que corresponde a 11,25% do total de registros. Porém, aqui temos o maior número de registros com a informação de *causa mortis*, computando 30 documentos. Consta, nesses documentos, que 16 sujeitos faleceram repentinamente; 03 faleceram devido à maligna, que Santos (2014) aponta como diarreia; 02 faleceram devido a problemas com hidropsia, *lexia* registrada nos documentos como *hydropezia* e *hydropica*; 02 faleceram por

afogamento; 02 morreram por complicação no parto; 01 faleceu devido a *encaio*, um tipo de prisão de ventre, por longo tempo; 01 morreu em decorrência de uma facada; 01 faleceu por problemas recorrentes de uma febre; 01 em decorrência de lepra e 01 foi a óbito devido à retenção de urina.

Entre os crioulos livres, a faixa etária com maior índice de mortalidade masculina concentra-se em 1 a 18 meses, com o total de 0,61%, enquanto a feminina situa-se entre 18 meses a 12 anos, computando igualmente 0,61%. Sobre os crioulos livres, apenas um registro apresenta *causa mortis*, em que uma crioula de 40 anos faleceu repentinamente.

Ademais, inventariamos registros de cabra, angola e caboclo livres, sendo que cada um deles corresponde a apenas 0,15% do total de óbitos, nenhum deles com a causa da morte especificada. Na categoria “não especificado”, 1,22% dos óbitos inventariados referem-se a homens e mulheres livres, sendo apenas uma morte identificada como complicações no parto.

## 5. Palavras finais

Este estudo teve como objetivo analisar, linguística e historicamente, um livro de registro de óbitos da Freguesia de Nossa Senhora da Penha de Corumbá, com documentos exarados entre os anos de 1847 e 1855 pelo Vigário Manoel Innocencio da Costa Campos. Para tanto, realizamos um inventário composto por informações referentes a cada falecido, baseando-nos no modelo proposto por Santos e Paula (2014), chegando ao total de 658 registros.

Partindo deste *corpus*, nossa análise fundamentou-se em dados como (i) o número de óbitos registrado por ano, em que estabelecemos uma média de 82 mortes ao ano; (ii) o número de óbitos de escravos, libertos e livres, que nos revelou um baixo número de registros de óbitos de escravos em Goiás (na região e nos anos estudados), o que pode revelar um possível remanejamento dos escravos, anteriormente empregados na exploração de minérios para novos locais de trabalho; (iii) o número de óbitos de acordo com as características, com destaque para a expressiva quantidade de registros de pardos; (iv) os locais de enterro, posto que era direito que todo cristão escolhesse o local de sua sepultura, ainda que isso não ocorresse de fato; (v) os sacramentos, informação pouco contida nos registros; (vi) as idades dos falecidos, seção em que apresentamos as faixas etárias com seus números de mortes registrados.

A análise dos dados arrolados no inventário nos possibilitou relacionar as informações contidas no livro de registros com aspectos da cultura e

da história do período em questão. Constatamos que os tipos de registros variavam de acordo com a condição social e jurídica do sujeito falecido, sendo os de mulheres e homens livres muito mais completos do que os de mulheres e homens forros, libertos e escravizados. Destarte, inferimos que ser livre, uma condição de nascença desses sujeitos, se fez notar, também, na hora de sua morte, originando diferenças nos registros.

Destacamos que as alforrias e as liberdades conquistadas pelos ex-escravos não garantiam, de fato, uma nova condição social, pois estes eram constantemente associados ao seu passado de escravizado e seus registros apresentavam a expressão “escravo que foi de”. Do ponto de vista do registro de óbito, o tratamento que recebiam continuava sendo o mesmo dos escravos, ainda que, em vida, gozassem da liberdade adquirida. Assim, na morte e na vida de outrora escravizado sua condição pouco diferia e o registro de morte é o signo em que, neste estudo, se inscrevem as memórias da escravidão.

Ao realizar este estudo, satisfizemos a função transcendente da Filologia (Spina 1977), pois apresentamos e analisamos dados importantes da língua, da história e da cultura do período oitocentista brasileiro, a partir dos 658 documentos em questão. Esperamos que esse estudo, que partiu de um *corpus* digitalizado de documentação histórica na fonte, além de trazer à baila possibilidades de investigação nas humanidades digitais, chancela contribuições para novas pesquisas dedicadas a essa temática, uma vez que ainda há uma vasta quantidade de documentos manuscritos que registram este momento da história brasileira carecendo de acesso para conhecimento, edição, descrição e investigação nas várias áreas do conhecimento em tais registros podem (e devem) ser estudados.

Ao fim e ao cabo, porque a história se faz múltipla e diversa conforme os signos que a escrevem e a inscrevem nas tramas da vida social, como procuramos demonstrar neste estudo, a demanda pela sua compreensão não pode prescindir de perscrutar no campo das humanidades, digitais ou manuscritas, suas fontes, suas memórias, suas linguagens, suas explicações possíveis.

## Referências

- Abbade, C. M. S. (2008). Filologia e o estudo do léxico. In J. S. Magalhães, & L. C. Travaglia (Eds.). *Múltiplas Perspectivas em Lingüística*, 1, 716–721. Uberlândia: EDUFU.
- Almeida, M. A. R., Amorim, A. M. & De Paula, M. H. (2017). Um cabra de cor ou um cabra da mãe: dinâmicas de sentido para “cabra” entre os séculos XVI e XIX. *Filologia e Lingüística Portuguesa (Online)*, 19, 143–161.

- Almeida, M. A. R. (2017). *Nas trilhas dos manuscritos: estudo lexical sobre a escravidão negra em Catalão-GO (1861–1887)* (Dissertação de mestrado, Universidade Federal de Goiás).
- Almeida, M. A. R., Amorim, A. M., Vaz, V. A. S. S. & De Paula, M. H. (2017). Crioulo, mulato e pardo: análise lexical das qualificações aos negros no Brasil oitocentista. In M. H. De Paula, M. P. Santos & S. M. Peres. (Eds.). *Perspectivas em estudos da linguagem*, 1, 159–170. São Paulo: Blücher.
- Araújo, R. M. (2011). *Preocupação com ‘bem morrer’ nas minas: análise de testamentos das mulheres de Vila do Carmo e seu termo, 1715–1750*. São Paulo: ANPUH SP.
- Barros, J. D'A. (2014). *A construção social da cor: diferença e desigualdades na formação da sociedade brasileira*. Petrópolis/RJ: Vozes.
- Bellotto, H. L. (2002). *Como fazer análise diplomática e análise tipológica de documentos de arquivo*. São Paulo: Arquivo do Estado, Imprensa Oficial do Estado.
- Bluteau, R. (1712–1728). *Vocabulário portuguez & latino*. Coimbra: Collegio das Artes da Companhia de Jesus. Disponível em: <http://dicionarios.bbm.usp.br/en/dicionario/edicao/1>. Consultado em: 15 jan. 2018
- Campos, M. I. C. (2011). *Livro de óbitos 1847 a 1867 nº 3*. Arquivo Histórico Estadual de Goiás.
- De Paula, M. H. (2007). *Rastros de velhos falares: léxico e cultura no vernáculo catalano* (Tese de doutoramento, Universidade Estadual Paulista Júlio de Mesquita Filho). Disponível em: [https://www.researchgate.net/publication/281964842\\_Rastros\\_de\\_velhos\\_falares\\_lexico\\_e\\_cultura\\_no\\_vernaculo\\_catalano](https://www.researchgate.net/publication/281964842_Rastros_de_velhos_falares_lexico_e_cultura_no_vernaculo_catalano). Consultado em: 10 jan. 2015.
- De Paula, M. H. & Almeida, M. A. R. (2016). Entre arraiais, vilas, cidades, comarcas e províncias: terminologias das representações do espaço no sudeste goiano no século XIX. *Revista (Con)Textos Linguísticos (UFES)*, 10, 153–157.
- De Paula, M. H., & Amorim, A. M. (2016). Léxico e cultura: breve análise de documentos oitocentistas sobre a escravidão negra em Catalão. *Intersecções (Jundiá)* 9 (4), 132–151.
- Ferreira, A. B. H. (2004) *Dicionário Eletrônico Aurélio Século XXI*. Rio de Janeiro: Editora Positivo (Versão 5.0).
- IPHAN (Instituto do Patrimônio Histórico e Artístico Nacional). *Igreja Matriz de Nossa Senhora da Penha de França, em Goiás, vai ser reinaugurada*. Disponível em: <http://portal.iphan.gov.br/noticias/detalhes/260>. Consultado em: 14 jan. 2018.
- Libby, D. C. & Paiva, E. F. (2005). *A Escravidão no Brasil: Relações Sociais, Acordos e Conflitos*. (2ª ed.) São Paulo: Moderna.
- Malheiro, A. M. P. (2014). *A Escravidão no Brasil. Ensaio Histórico-Jurídico-Social*. São Paulo: Poeteiro Editor Digital.
- Nunes, V. M. M. (2008). *Glossário de termos sobre religiosidade*. Aracaju: Tribunal de Justiça; Arquivo Judiciário do Estado do Sergipe.
- Paiva, E. F. (2014). *Dar nome ao novo: uma história lexical das Américas portuguesa e espanhola, entre os séculos XVI e XVIII (as dinâmicas de mestiçagem e o mundo do*

- trabalho*) (Tese para concurso de Professor Titular de História do Brasil, Universidade Federal de Minas Gerais).
- Salles, G. V. F. (1992). *Economia e escravidão na capitania de Goiás*. Goiânia: Centro Editorial e Gráfico da UFG.
- Santos, J. C. (2014). A hora derradeira de homens e mulheres africanos e seus descendentes: alguns apontamentos sobre os óbitos, Santo Amaro, Sergipe, 1802–1835. *Revista do IHGSE*, 44, 339–364.
- Santos, W. F. & Paula, J. H. (2014). Escravidão em Goiás: mortalidade branca e escrava na Vila de Santa Luzia entre os anos de 1786 a 1814. *8º Seminário de iniciação científica e tecnológica*, 8, Itumbiara-GO.
- Silva Pinto, L. M. (1832). *Diccionario da Lingua Brasileira*. Disponível em: <http://dicionarios.bbm.usp.br/en/dicionario/edicao/3>. Consultado em: 15 jan. 2018.
- Spina, S. (1977). *Introdução à Edótica: crítica textual*. São Paulo: Cultrix (Editora da Universidade de São Paulo).
- Vide, D. S. M. (1853). *Constituições Primeiras do Arcebispado da Bahia*. Typographia São Paulo: Senado Federal.
- Xavier, V. R. D. (2010). Administração ou escravização indígena? O que dizem os documentos coloniais goianos. *Signotica (UFG)*, 22, 465–478.
- Xavier, V. R. D. (2012). *Conexões léxico-culturais sobre as minas goianas setecentistas no Livro para servir no registro do Caminho Novo de Parati* (Tese de doutoramento, Universidade de São Paulo). Disponível em: <http://www.teses.usp.br/teses/disponiveis/8/8142/tde-29082012-100504/pt-br.php>. Consultado em: 15 nov. 2017.

[recebido em 31 de março de 2018 e aceite para publicação em 11 de novembro de 2018]

## Corpora nas humanidades digitais

Identidade e diferenças na terminologia da *fauna* e da *flora*: notas sobre um estudo comparativo entre as línguas portuguesa, inglesa, italiana e espanhola  
Sabrina de Cássia Martins

*Agroquímico, biocida, pesticida, plaguicida e producto fitosanitário*: uma pesquisa com *corpus*  
Mauren Thiemy Ito Cereser;  
Cleci Regina Bevilacqua

Quando o léxico dá bandeira – aspectos cognitivo-discursivos da mudança semântica na construção de brasileirismos em registros lexicográficos luso-brasileiros  
Anderson Salvaterra Magalhães;  
Janderson Lemos de Souza

Características identificadoras e dificuldades na aplicação de listas para a anotação de entidades geográficas mencionadas  
Afonso Xavier Canosa

Uma versão em português europeu do *C-test*  
Masayuki Yamada

Aplicação de ferramentas para coleta e análise de dados em linguística  
Roberlei Alves Bertucci

Análise diacrónica dos tempos compostos *tinha feito, terei feito e teria feito* na língua portuguesa  
Jan Hricsina

Análise contrastiva das formas de tratamento ao interlocutor no teatro brasileiro e português dos séculos XIX e XX  
Ana Carolina Morito Machado

The role of pragmatic markers in academic spoken Interlanguage: a corpus-based study of a group of Brazilian EFL university students  
Bárbara Malveira Orfanò; Ana Larissa Adorno Marciotto Oliveira;  
Spencer Barbosa da Silva

Corpus Stylistics in translation-oriented text analysis: Approaching the work of Denton Welch from a Functionalist perspective  
Guilherme da Silva Braga

Em vida e na hora da morte também: o que dizem registros de óbito oitocentistas da freguesia de Nossa Senhora da Penha de Corumbá (1847–1855)  
Maria Helena de Paula;  
Amanda Moreira de Amorim