UNIVERSIDADE FEDERAL DE UBERLÂNDIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

DOUTORADO EM ESTUDOS LINGUÍSTICOS

MARIA VIRGÍNIA DIAS DE ÁVILA

DESCRIÇÃO ETIMOLÓGICA DO LÉXICO INDIANISTA EM JOSÉ DE ALENCAR: UMA ANÁLISE LEXICOGRÁFICA DIRECIONADA POR *CORPUS*

UBERLÂNDIA - MG

2018

MARIA VIRGÍNIA DIAS DE ÁVILA

DESCRIÇÃO ETIMOLÓGICA DO LÉXICO INDIANISTA EM JOSÉ DE ALENCAR: UMA ANÁLISE LEXICOGRÁFICA DIRECIONADA POR *CORPUS*

Tese apresentada à banca examinadora do Programa de Pós-Graduação em Estudos Linguísticos – Curso de Mestrado e Doutorado – do Instituto de Letras e Linguística da Universidade Federal de Uberlândia, como requisito parcial para obtenção do título de Doutora em Estudos Linguísticos.

Área de concentração: Estudos em Linguística e Linguística Aplicada.

Linha de pesquisa: Teoria, Descrição e Análise Linguística

Orientador: Prof. Dr. Ariel Novodvorski

UBERLÂNDIA - MG

2018

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

A958d
2018

Ávila, Maria Virgínia Dias de, 1968-
Descrição etimológica do léxico indianista em José de Alencar [recurso eletrônico] : uma análise lexicográfica direcionada por corpus / Maria Virgínia Dias de Ávila. - 2018.

Orientador: Ariel Novodvorski.
Tese (doutorado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Estudos Linguísticos.
Modo de acesso: Internet.
Disponível em: http://dx.doi.org/10.14393/ufu.te.2018.625
Inclui bibliografia.
Inclui ilustrações.

1. Linguística. 2. Língua portuguesa - Etimologia. 3. Linguística de corpus. 4. Alencar, José de, 1829-1877 - Crítica e interpretação. I. Novodvorski, Ariel (Orient.) II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Estudos Linguísticos. III. Título.

CDU: 801

Gerlaine Araújo Silva - CRB-6/1408

MARIA VIRGÍNIA DIAS DE ÁVILA

DESCRIÇÃO ETIMOLÓGICA DO LÉXICO INDIANISTA EM JOSÉ DE ALENCAR: UMA ANÁLISE LEXICOGRÁFICA DIRECIONADA POR *CORPUS*

Tese apresentada à banca examinadora, para a obtenção do título de Doutora em Estudos Linguísticos no Programa de Pós-Graduação em Estudos Linguísticos – Curso de Mestrado e Doutorado – do Instituto de Letras e Linguística - Universidade Federal de Uberlândia.

Uberlândia, 09 de julho de 2018.

Banca Examinadora

_____
Prof. Dr. Ariel Novodvorski – ILEEL/PPGEL - UFU
Orientador

_____
Profa. Dra. Kênia Maria de Almeida Pereira – PPLET - UFU

_____
Profa. Dra. Eliana Dias – PROFLETRAS - UFU

_____
Prof. Dr. Evandro Silva Martins – ILEEL – UFU

_____
Prof. Dr. Guilherme Fromm – ILEEL/PPGEL - UFU

Aos meus filhos, Maiko e Lucas...
Luzes do meu caminhar!

AGRADECIMENTOS

A Deus, por ter corpo e alma sãos.

Ao Prof. Dr. Ariel Novodvorski, que aceitou o desafio de me orientar em tão pouco tempo para escrever uma tese. Obrigada pela serenidade, leveza e sabedoria com que conduziu a orientação, demonstrando a confiança necessária para eu realizar a pesquisa e finalizar o curso. Mostrou-me que o ato da pesquisa pode ser feliz. Com muita admiração, serei sempre grata.

Aos meus filhos, Maiko e Lucas, pelo amor incondicional. Por eles, também, meu esforço.

Aos meus pais, pelas palavras de doçura a cada encontro e pelas mesmas palavras afetuosas, quando minhas visitas foram impossibilitadas pela reclusão e dedicação ao trabalho na pesquisa.

A todos da minha família, por me apoiarem e compreenderem minhas ausências.

Ao meu companheiro de vida, pelas doses generosas e diárias de afeto e de apoio, todos os apoios, tornando meus dias mais fáceis e felizes.

À Fernanda Borges, pela presença carinhosa em nossas vidas, especialmente na vida de meu filho.

À Profa. Dilma Mello, pelo empenho para que eu pudesse participar do Processo Seletivo e pelo aprendizado e confiança nos primeiros anos de orientação. Agradeço, também, pela orientação do trabalho de Área Complementar.

À Lorena e ao Fernando, por assumirem as atividades do Processo Seletivo quando eu não pude fazê-lo, por participar do processo.

À Luana, por assumir os trabalhos do PPGEL, durante a minha licença para a pesquisa.

Aos amigos do GPNEP, Grupo de Pesquisa Narrativa e Educação de Professores, pelas interlocuções proveitosas durante a primeira fase do meu processo de doutoramento.

Aos amigos do grupo de pesquisa GECon - Grupo de pesquisa em Estudos Contrastivos- por compartilhar conhecimentos e experiências de pesquisa.

Ao Prof. Dr. Waldenor Barros Moraes Filho, pela frase despretensiosa de corredor que me fez perceber que era hora de iniciar o Doutorado.

À Profa. Dra. Valeska Virgínia Soares Sousa, pelas primeiras observações, muito valiosas, sobre o meu projeto de pesquisa e pelo gesto delicado de cuidado ao me auxiliar com os textos escritos em língua inglesa.

Aos Professores Doutores Guilherme Fromm e Evandro Silva Martins, pelas valiosas contribuições durante a banca de qualificação da tese.

Aos professores Doutores Evandro Silva Martins, Kênia Maria de Almeida Pereira, Eliana Dias e Guilherme Fromm pela leitura cuidadosa do meu trabalho para a defesa da tese e pelas ponderações e sugestões de aperfeiçoamento durante a defesa. Muito obrigada pelo te m dedicado a meu trabalho e a mim.

Às Professoras Doutoras Valeska Virgínia Soares Souza e Emeli Borges Pereira Luz, pelas orientações durante a qualificação do Projeto de pesquisa.

Aos amigos do GPNEP, Gyzely Lima, Lauro e Geralda, pelo apoio na leitura dos textos escritos em língua inglesa.

À Liana Castro e à Gyzely Lima, pelo auxílio na tradução do resumo da tese para as línguas espanhola e inglesa. Obrigada, amigas queridas!

Às queridas amigas, Regina Nascimento e Anair Valênia, pela leitura e observações na primeira versão do meu projeto de pesquisa.

Ao Antonio, por dividir comigo os cuidados com o Lucas, o que me proporcionou tranquilidade e segurança para estar longe, quando me foi exigido.

Ao Prof. Dr. Cleudemar Alves Fernandes, por me mostrar novos horizontes e por ter me encorajado, a cada dia.

À Profa. Dra. Dilma Mello e ao Prof. Dr. Cleudemar Alves Fernandes, pelo apoio na coordenação do PPGEL, com a flexibilização dos meus horários de trabalho.

Aos meus alunos, por participarem do meu processo de doutoramento e me proporcionarem o sentimento da certeza da escolha da profissão.

Aos meus coordenadores Cléria e Tiago, por facilitarem, de todas as formas, o meu trabalho.

À Universidade Federal de Uberlândia, pela licença concedida.

Por fim, agradeço a mim mesma, não por pretensão ou autossuficiência, mas por reconhecer o meu próprio mérito no percurso da pesquisa e escrita desta tese. Trabalhei intensamente, enfrentei os problemas do dia a dia com determinação, não desisti quando a pressão parecia insuportável, abdiquei de muitas companhias e muitos momentos de prazer. Enfim, com orgulho de mim mesma, sinto-me feliz e realizada.

Tudo passa sobre a terra!
Alencar.

RESUMO

A presente tese tem como objetivo desenvolver um estudo com base no léxico indianista de José de Alencar em suas obras consideradas indianistas: *Iracema*, *O Guarani* e *Ubirajara*. Nosso estudo abrangeu dois temas gerais: por um lado, verificar se Alencar utiliza um léxico específico em suas obras indianistas; por outro, analisar o léxico indianista baseando-nos na perspectiva da Etimologia Ficcional Contextual. E, como terceiro objetivo, propor a elaboração de um Vocabulário eletrônico dos vocábulos indianistas, a ser disponibilizado *online*. Em nossa proposta, dada a diversidade, articulamos saberes relacionados ao Léxico, naquilo que se refere à Lexicologia e à Lexicografia; abordamos questões sobre a caracterização e constituição de obras lexicográficas e sobre os dicionários eletrônicos, em específico. Tendo em vista os procedimentos de organização do léxico, também discorremos sobre campos semânticos. Para dar conta da nossa proposta de análise pelo viés da Etimologia, expandimos o conceito para Etimologia Ficcional Contextual e, por fim, nosso olhar para o arcabouço teórico-metodológico centrou-se nos pressupostos da Linguística de *Corpus*, concebendo-a como abordagem, principalmente, pelo viés das pesquisas guiadas por *corpus*. Para esta pesquisa, compilamos nosso *corpus* de estudos com as obras indianistas de Alencar e, para contrastar e verificar se o autor utiliza um léxico específico nas obras indianistas, compilamos o *corpus* de referência com as demais obras do mesmo autor. Para processamento dos *corpora*, recorremos ao programa *WordSmith Tools* 6.0 e suas três ferramentas: *WordList*, *KeyWords* e *Concord*, que nos permitiram a identificação dos vocábulos indianistas, acesso aos contextos abonatórios das ocorrências, além do contraste entre o *corpus* de estudo e os *corpora* de referência. Para a análise etimológica, recorremos também ao *Corpus do Português* (DAVIES, 2016) e aos dicionários de consulta e exclusão. A partir dos dados, elaboramos 17 fichas lexicográficas que serviram de amostra para o trabalho de organização do Vocabulário indianista de Alencar. Concluímos que Alencar utiliza um léxico específico em suas obras indianistas, considerando as palavras que são chave no *corpus* de estudo em relação aos *corpora* de referência, como também pelo fato de, aproximadamente, 41% de vocábulos substantivos e adjetivos indígenas utilizados pelo autor em suas obras, serem indianistas. Além disso, verificamos também que o autor cria étimos em seus romances para nomear, conforme seus desejos literários e linguísticos. Alencar utiliza o critério de segmentação de um vocábulo com base em conhecimentos de língua indígena, para, com os fragmentos, criar novos vocábulos de acordo com o contexto ficcional no interior dos romances. Por fim, consideramos a proposta do Vocabulário com os vocábulos indígenas, a ser disponibilizado *online*, de grande importância não apenas para os consulentes leitores de Alencar, como também para aqueles que desejarem empreender estudos sobre o vocabulário indígena, que contribuiu para a formação da língua portuguesa brasileira.

Palavras-chave: Etimologia Ficcional Contextual. Indianismo em Alencar. Linguística de *Corpus*. Vocabulário indianista. Léxico indianista.

ABSTRACT

This thesis aims at developing a study based on specific lexicon features used by José de Alencar, that is, indigenous lexicon, in his considered indianist works, such as: *Iracema*, O *Guarani* and *Ubirajara*. This study is comprised of two general topics: verifying if the author uses a specific lexicon in his indianist works, and analyzing such lexicon from the perspective of Contextual Fiction Etymology. In addition, as a third objective, we propose the creation of an electronic vocabulary of such indianist terms making them available online. Due to the diversity, in our proposal, we articulated knowledge related to lexicon, concerning Lexicology and Lexicography and we approached issues relating to the characterization and constitution of lexicography words, specifically, electronic dictionaries. Considering lexicon organization procedures, we also discussed semantic fields. In order to carry out our analysis through the Etymology scope, we expanded the concept of Contextual Fiction Etymology and, lastly, our approach on the theoretical and methodology support was centered on a Corpus Linguistics basis approach, mainly in a corpus-guided research scope. We compiled our studied corpus with Alencar's indianist works, and, to contrast and verify if the author uses a specific lexicon in such works, we compiled a reference corpus with other works of the same author. Corpora processing was done in *WordSmith Tools* 6.0 and its three main tools: *WordList*, *KeyWords* and *Concord*, which enabled us identify indianist terms, access abonatory contents occurrences, as well as the contrast between the studied and reference corpus. For the etymology analysis, we used a *corpus* of the *Portuguese language* (DAVIES, 2016), and also regular and exclusion dictionaries. From the data, we elaborated 17 lexicography forms that served as a sample for the organization of Alencar's indianist vocabulary. It is our conclusion that Alencar uses a specific lexicon in his indianist works, considering the keywords in the studied corpus in relation to the reference corpora, also due to the fact that approximately 41% of the noun and adjectives used in his works are indigenous. In addition, we also verified that the author creates etymons in his novels to nominate, according to his literary and linguistics wishes. The author uses the segmentation criterion of a term based on indigenous language knowledge to, from the fragments, create new terms according to the fiction context of his novels. Lastly, we consider of great importance that the Indigenous Vocabulary proposal is available online, not only for Alencar's readers, but also for those that wish to undertake studies on indigenous vocabulary, which contributes for Brazilian Portuguese formation.

Keywords: Contextual Fiction Etymology. Indianism in Alencar. Corpus linguistics. Indianist vocabulary. Indianist lexicon.

RESUMEN

Esta tesis tiene como objetivo desarrollar un estudio con base en el léxico indianista de José de Alencar en sus obras consideradas indianistas: *Iracema*, *O Guarani* y *Ubirajara*. Nuestro estudio abarcó dos temas generales: por un lado, verificar si Alencar utiliza un léxico específico en sus obras indianistas; por otro, analizar el léxico indianista basándonos en la perspectiva de la Etimología Ficcional Contextual. Y, como tercer objetivo, proponer la elaboración de un Vocabulario electrónico de los vocablos indianistas, que se harán disponibles *online*. En nuestra propuesta, dada la diversidad, articulamos saberes relacionados al Léxico, en lo que se refiere a la Lexicología y a la Lexicografía; abordamos cuestiones sobre la caracterización y constitución de obras lexicográficas y sobre los diccionarios electrónicos, en específico. Teniendo en cuentra los procedimientos de organización del léxico, también discurrimos sobre campos semánticos. Para abarcar nuestra propuesta de análisis por la línea de la Etimología, expandimos el concepto para Etimología Ficcional Contextual y, por fin, nuestra mirada para hacia el eje teórico-metodológico se centró en los presupuestos de la Lingüística de *Corpus*, concibiéndola como abordaje, principalmente, por la línea de las investigaciones guiadas por *corpus*. Para esta pesquisa, compilamos nuestro *corpus* de estudios con las obras indianistas de Alencar y, para contrastar y verificar si el autor utiliza un léxico específico en las obras indianistas, compilamos un *corpus* de referencia con las demás obras del mismo autor. Para procesamiento de los *corpus*, recurrimos al programa *WordSmith Tools* 6.0 y sus tres herramientas: *WordList*, *KeyWords* y *Concord*, que nos permitieron la identificación de los vocablos indianistas, acceso a los contextos de cita de las ocurrencias, además del contraste entre el *corpus* de estudio y los *corpus* de referencia. Para el análisis etimológico, recurrimos también al *Corpus do Portugués* (DAVIES, 2016), como *corpus* de consulta, y a los diccionarios de consulta y exclusión. A partir de los datos, elaboramos 17 fichas lexicográficas que sirvieron de muestra para el trabajo de organización del Vocabulario indianista de Alencar. Concluimos que Alencar utiliza un léxico específico en sus obras indianistas, considerando las palabras que son clave en el *corpus* de estudio, en relación con los *corpus* de referencia*, y* también por el hecho de que, aproximadamente, un 41% de vocablos sustantivos y adjetivos indígenas utilizados por el autor en sus obras son indianistas. Más allá de todo eso, verificamos también que el autor crea étimos en sus novelas con el objetivo de nombrar personajes y cosas, según sus deseos literarios y lingüísticos. Alencar utiliza el criterio de segmentación de un vocablo con base en conocimientos de lengua indígena; posteriormente, con los fragmentos ya segmentados, crea nuevos vocablos de acuerdo con el contexto ficcional en el interior de las novelas. Por fin, consideramos la propuesta del Vocabulario con los vocablos indígenas, que se harán disponibles *online*, de gran importancia no sólo para los consultores y/o lectores de Alencar, sino también para aquellos que deseen emprender estudios sobre el vocabulario indígena, que contribuyó sobremanera para la formación de la lengua portuguesa brasileña.

Palabras clave: Etimología Ficcional Contextual. Indianismo en Alencar. Lingüística de *Corpus*. Vocabulario indianista. Léxico indianista.

LISTA DE ILUSTRAÇÕES

Figura 1: Detalhe da tela inicial do *WST* com os nomes das ferramentas 91

Figura 2: *Corpus* de estudo..... 100

Figura 3: *Corpus* de Referência..... 101

Figura 4: Dados referentes a palavra **anajê** no *Corpus do Português*.....105

Figura 5: Vista parcial do armazenamento dos *corpora* de José de Alencar.....109

Figura 6: Procedimento de normalização do *corpus* 110

Figura 7: Procedimento de revisão do texto 111

Figura 8: Elementos a serem removidos dos *corpora* de estudo e referência 112

Figura 9: Elementos a serem removidos dos *corpora* de estudo e referência 112

Figura 10: Lista das 17 primeiras palavras do *corpus* de estudo 115

Figura 11: Lista das 17 primeiras palavras do corpus de estudo 115

Figura 12: Palavras-chave com CorpRef-Alencar..... 117

Figura 13: Palavras-chave com o CorpRef-Lácio-Web.....117

Figura 14: Palavras-chave com o CorpRef-AcadTeses 118

Figura 15: Palavras-chave com o CorpRef-AcadTeses. 118

Figura 16: Visão parcial da lista de palavras-chave: *corpus* de estudo com CorpRef-Alencar.....120

Figura 17: Visão parcial da lista de palavras-chave: *corpus* de estudo com CorpRef-Lácio-Web..... 120

Figura 18: Vista parcial do procedimento de extração do léxico indianista..... 122

Figura 19: Candidatos a vocábulos indianistas.....122

Figura 20: Vista parcial das palavras não indígenas excluídas da lista de palavras 123

Figura 21: Lista de candidatos a vocábulos após consulta ao *Houaiss* 125

Figura 22: Linhas de concordância da palavra **Peri**..... 128

Figura 23: Lista do léxico indianista de José de Alencar 129

Figura 24: Listas das linhas de concordância do vocábulo **atiati**.....132

Figura 25: Lista das linhas de concordâncias do vocábulo **jatobá**.....132

Figura 26: Palavras-chave do *corpus* de estudo em contraste com o *corpus* de referência CorpRef-Alencar 141

Figura 27: Vista do trecho do romance *As Minas de Prata* em que Alencar emprega o vocábulo **Peri** 142

Figura 28: Contraste com o *corpus* de referência CorpRef-AcadTeses 144

Figura 29: Contraste com o *corpus* de referência CorpRef-Lácio-Web..... 144

Figura 30: Contraste com o *corpus* de referência CorpRef-Nov 145

Figura 31: Corpus de estudo em contraste com o corpus de referência CorpRef-Alencar....148

Figura 32: *Corpus* de estudo em contraste com o *corpus* de referência CorpRef-AcadTeses 148

Figura 33: *Corpus* de estudo em contraste com o *corpus* de referência CorpRef-Lácio-Web 149

Figura 34: Corpus de estudo em contraste com o corpus de referência CorpRef-Nov. 149

Figura 35: Vista parcial das 20 primeiras palavras indianistas do *corpus* de estudo 154

Figura 36: Linhas de concordância da palavra **Araquém**..... 157

Figura 37: Linhas de concordância da palavra **aroeira**..... 158

Figura 38: Linhas de concordância do vocábulo **Iracema**..... 167

Figura 39: Vocábulo **Iracema** em contexto ampliado no texto *Eu* de Augusto dos Anjos ...168

Figura 40: Vocábulo **Iracema** em contexto ampliado no livro *O Guarani* de José de Alencar 168

Figura 41: Linhas de concordância do vocábulo **Moacir** 171

Figura 42: Linhas de concordância do vocábulo **Moacir** 172

Figura 43: Linhas de concordância do vocábulo **Irapuã** 173

Figura 44: Linhas de concordância do vocábulo **Coatiabo** 174

Figura 45: Linhas de concordância do vocábulo **Maranguab**..... 175

Figura 46: Linhas de concordância do vocábulo **Maranguab**..... 176

Figura 47: Linhas de concordância do vocábulo **jacarecanga** 177

Figura 48: Linhas de concordância do vocábulo **maracatim** 179

Figura 49: Linhas de concordância do vocábulo **igara** 180

Figura 50: Linhas de concordância do vocábulo **abaeté**..... 181

Figura 51: Linhas de concordância do vocábulo **anhangá** 182

Figura 52: Linhas de concordância do vocábulo **anhangá** 182

Figura 53: Linhas de concordância do vocábulo **anhanga** 183

Figura 54: Linhas de concordância do vocábulo **moquém** 184

Figura 55: Linhas de concordância do vocábulo **moquém** 185

Figura 56: Contexto ampliado do vocábulo **moquém**..... 185

Figura 57: Linhas de concordância do vocábulo **carimã** 186

Figura 58: Linhas de concordância do vocábulo **carimã** 186

Figura 59: Linhas de concordância do vocábulo **camucim** 188

Figura 60: Linhas de concordância do vocábulo **camucim** 189

Figura 61: Linhas de concordância do vocábulo **camuci**..... 189

Figura 62: Linhas de concordância do vocábulo **inúbia** 190

Figura 63: Linhas de concordância do vocábulo **guará** 191

Figura 64: Linhas de concordância do vocábulo **gará** 192

Figura 65: Lista de concordância do vocábulo **andira** no *corpus* de estudo..... 211

Figura 66: Linhas de concordância do vocábulo **jatobá** 220

Figura 67: Vista da página inicial do Vocabulário *online*..... 226

Figura 68: Vista do Vocabulário a partir da escolha de ícones 227

Figura 69: Vista dos campos semânticos..... 231

Figura 70: Dados sobre o vocábulo **jatobá** a partir da escolha do consulente 232

Figura 71: Vista da tela do computador a partir da consulta pelas palavras..... 233

Imagem 1: José de Alencar..... 34

Imagem 2: Capa e folho de rosto do livro *Iracema* (1ª edição) e capa de *O Guarani* (1ª edição) e capa da 3ª edição de *Ubirajara*..... 42

Imagem 3: Folha do Caderno X dos *Apontamentos Diversos* de José de Alencar 164

Gráfico 1: Identificação da quantidade de palavras indianistas nas obras de Alencar 160

Quadro 1: Tipologias de obras lexicográficas com base em Barbosa 58

Quadro 2: Modelo de verbete proposto por Barbosa..... 64

Quadro 3: Características dos dicionários impressos e eletrônicos 67

Quadro 4: *Corpus* de estudo 99

Quadro 5: *Corpus* de Referência CorpRef-Alencar 101

Quadro 6: Sistematização dos dicionários..... 126

Quadro 7: Léxico indianista distribuído em campos semânticos 130

Quadro 8: Quantidade de palavras com frequência zero em relação ao total de itens nos *corpora* de referência..... 150

Quadro 9: Lista das 367 palavras consideradas indígenas do *corpus* de estudo..... 155

Quadro 10: Palavras consideradas indígenas que não constam dos dicionários de exclusão 159

Quadro 11: Distribuição dos vocábulos indígenas por campo semântico 228

LISTA DE TABELAS

Tabela 1: Classificação do *corpus* em relação à quantidade de palavras 85

Tabela 2: *Corpus* de estudo e *corpus* de referência: o autor com ele mesmo 95

Tabela 3: Comparação entre o *corpus* de estudo e os *corpora* de referência 98

LISTA DE ABREVIATURAS E SIGLAS

AGC – *Dicionário Histórico das palavras Portuguesas de origem Tupi*, de Antônio Geraldo da Cunha (1998)

AMS - *Diccionário da Lingua Portugueza*, de Antonio de Moraes Silva (1789)

GD - *Dicionário de língua Tupy*, de A. Gonçalves Dias (1858)

JA – José de Alencar

LC – Linguística de *Corpus*

LCT – *O Dicionário Tupi-português*, de Luiz Caldas Tibiriçá (1984)

LMSP - *Dicionário da Lingua Brasileir*a, de Luiz Maria da Silva Pinto (1832)

*KW*S - *KeyWords*

RB - *Diccionário da Lingua Portugueza*, de Raphael Bluteau (1789)

TS - *O tupi da geografia nacional*, de Theodoro Sampaio (1901)

UFU – Universidade Federal de Uberlândia

WST – *WordSmith Tools*

SUMÁRIO

1 INTRODUÇÃO 20

1.1 José de Alencar: alguns estudos 23

1.2 Justificativas da pesquisa e da tese 27

1.3 Hipótese e questões de pesquisa 29

1.4 Objetivos da pesquisa 30

***1.4.1 Objetivo geral* 30**

***1.4.2 Objetivos específicos* 30**

1.5 Organização da tese 30

CAPÍTULO 2 - UNIVERSO DA PESQUISA: ASPECTOS HISTÓRICOS 33

2.1 José de Alencar: breve biografia, o nacionalismo e o indianismo 33

***2.1.1 José de Alencar: breve biografia* 34**

***2.1.2 José de Alencar: o nacionalista* 36**

***2.1.3 José de Alencar e o indianismo* 39**

2.2 A tríade indianista de José de Alencar: *O Guarani, Iracema* e *Ubirajara* 42

CAPÍTULO 3 - FUNDAMENTOS TEÓRICOS 47

3.1 Léxico 47

3.2 Lexicologia e Lexicografia 52

3.3 Caracterização de obras lexicográficas 54

3.4 Constituição de obras lexicográficas 59

***3.4.1 Macroestrutura* 61**

***3.4.2 Microestrutura* 62**

3.5 Dicionários eletrônicos 66

3.6 Campos semânticos 70

3.7 Antropônimos e topônimos: algumas considerações 73

3.8 Etimologia 75

***3.8.1 Etimologia Ficcional Contextual* 80**

3.9 Linguística de *Corpus*, *corpus*, chavicidade e ferramentas do *WST*: definições e características 83

***3.9.1 Linguística de* Corpus *e* Corpus 83**

***3.9.2 Chavicidade* 89**

***3.9.3 Ferramentas de análise lexical:* WordList, KeyWords *e* Concord 91**

CAPÍTULO 4 - *CORPUS* E METODOLOGIA 93

4.1 Corpus 93

4.1.1 **Corpus** ***de José de Alencar: o autor com ele mesmo*** **93**
4.1.2 **Corpus** ***de estudo e*** **corpora** ***de referência*** **96**
4.1.3 Perfil dos textos que compõem o **corpus** *de estudo* **98**
4.1.4 Perfil dos textos de José de Alencar que compõem o **corpus** *de referência* **100**
4.1.5 **Corpus** ***de consulta*** **103**

4.2 Procedimentos Metodológicos **105**
4.2.1 Levantamento e compilação do **corpus** *de estudo e* **corpus** *de referência de José de Alencar* **107**
4.2.2 Armazenamento dos **corpora** **108**
4.2.3 Preparação do **corpus** *de estudo* **109**
4.2.4 Extração das listas de palavras para o levantamento dos dados estatísticos **114**
4.2.5 Extração das listas de palavras-chave **116**
4.2.6 Análise contrastiva das diferentes listas de palavras, a partir dos diferentes **corpora** *de referência* **119**
4.2.7 Identificação do léxico indianista **121**
4.2.8 Agrupamento do léxico indianista por campos semânticos **130**
4.2.9 Extração e análise das linhas de concordâncias com o léxico indianista **131**
4.2.10 Utilização de dicionários e de **corpus** *de consulta, na análise e descrição etimológica do léxico de Alencar* **133**
4.2.11 Elaboração das fichas lexicográficas com o léxico indianista identificado e analisado **135**
4.2.12 Elaboração de proposta de verbete **138**

CAPÍTULO 5 - ANÁLISE CONTRASTIVA DO *CORPUS* DE ESTUDO DE ALENCAR COM OS *CORPORA* DE REFERÊNCIA **140**

5.1 Análise das palavras-chave do *corpus* de estudo **141**

5.2 *Corpus* de estudo em contraste com os *corpora* de referência: análise das palavras de frequência zero **147**

5.3 Vocábulos indígenas do *corpus* de estudo **152**

CAPÍTULO 6 - A ETIMOLOGIA FICCIONAL CONTEXTUAL EM ALENCAR: A CRIAÇÃO DE VOCÁBULOS INDÍGENAS **163**

CAPÍTULO 7 - LÉXICO INDIANISTA DE JOSÉ DE ALENCAR: PROPOSTA PARA UM VOCABULÁRIO *ONLINE* **195**

7.1 Elaboração do vocabulário indianista **195**

7.2 Fichas lexicográficas e verbetes: um recorte do léxico indianista **197**

7.3 Proposta para elaboração de Vocabulário *online* do léxico indianista de José de Alencar **223**

CONCLUSÕES E CONSIDERAÇÕES **235**

REFERÊNCIAS **241**

APÊNDICES **249**

APÊNDICE A – Quadro comparativo entre as 20 primeiras palavras extraídas por meio da ferramenta KeyWords, do WST. *Corpus* de Estudo com os *corpora* de referência. . 249

APÊNDICE B – Quadro comparativo entre as 20 primeiras palavras com frequência zero extraídas por meio da ferramenta KeyWords, do WST. *Corpus* de Estudo com os *corpora* de referência. 250

APÊNDICE C – Lista das 367 palavras com indicação das obras em que são utilizadas 252

1 INTRODUÇÃO

O meu gosto[1] pela literatura retrocede aos tempos de Ensino Fundamental, quando li, pela primeira vez, *O Guarani* e *Iracema* de José de Alencar, *A Moreninha* de Joaquim Manuel de Macedo, *A escrava Isaura* de Bernardo Guimarães, dentre outros. Ainda sem uma perspectiva crítica, mas com um olhar e sensações que adentravam as histórias, dançando nos bailes com os longos vestidos de festas ou torcendo pelo amor impossível entre índios e brancos. Era assim que lia os livros envelhecidos de que a biblioteca do colégio dispunha para empréstimo.

Com um certo gosto, arriscava também a escrever, embora a veia literária não me tenha sido generosa. O curso de Letras, então, foi a escolha na qual poderia conciliar o prazer pela Literatura, o gosto pela escrita e o desejo de ser professora. Desde o primeiro dia de aula na Graduação do curso de Letras, estava convicta de que seria professora de Literatura. E consegui. Fui professora de Literatura Infanto-juvenil por dez anos no Ensino Fundamental e de Literatura no Ensino Médio, em escolas particulares de minha cidade. Porém, com o decorrer dos anos, outras atividades relacionadas ao ensino e à aprendizagem de Língua Portuguesa também despertavam meu interesse e, de alguma forma, me chamavam a outras possibilidades.

Esse chamamento se fortaleceu quando fui aluna, no curso de Mestrado, do Prof. Dr. Evandro Silva Martins que ministrou disciplinas relacionadas à Morfologia e à Lexicologia e nos sugeriu alguns temas para os projetos, dentre eles "O símile em José de Alencar". Não tive dúvidas e o escolhi. Convicta da escolha, pois poderia associar a Literatura aos Estudos Linguísticos, iniciei as leituras sobre o tema.

A partir das leituras e das orientações do Prof. Dr. Evandro, alteramos o foco para "o léxico indianistas nas obras de José de Alencar", porém, dada a extensão das obras sobre essa temática, foi necessário um recorte. Então, o romance escolhido, para a pesquisa do Mestrado, foi *Iracema*. Ficara para o Doutorado a continuação deste estudo. Sempre tive um desejo de expandir o estudo sobre o léxico de Alencar e de divulgar um vocabulário dos vocábulos indianistas das obras do autor. Na dissertação "O léxico indianista em José de Alencar: uma análise parcelar", o glossário ficou restrito ao romance *Iracema* e tratamos apenas dos nomes

[1] A introdução será escrita em primeira pessoa, por trazer também antecedentes pessoais.

substantivos, porém, agora, nos propomos a expandir o estudo sobre o léxico indianista em outras perspectivas também.

Quando, então, iniciei a empreitada para a seleção para o Curso de Doutorado, resolvi enveredar por outros campos e, assim, me inscrevi para uma pesquisa na subárea "Ensino e Aprendizagem de Línguas". Os estudos transcorriam dentro da expectativa do curso, porém ao iniciar as orientações para o trabalho de área complementar[2], sob orientação do Prof. Dr. Ariel Novodvorski, o desejo de retornar aos estudos sobre Alencar e o léxico se sobrepôs a qualquer outra possibilidade de pesquisa. Então, a mudança da pesquisa me trouxe novamente ao léxico indianista de José de Alencar. Mais especificamente, iniciamos as orientações e as pesquisas que desembocaram nesta tese, no início do mês de abril de 2017.

Dessa maneira, o tema desta pesquisa se insere no âmbito do grupo de pesquisa *GECon - Grupo de pesquisa em Estudos Contrastivos*-, coordenado pelo Prof. Dr. Ariel Novodvorski, da linha de pesquisa Teoria, descrição e análise linguística do Programa de Pós-Graduação em Estudos Linguísticos da Universidade Federal de Uberlândia (UFU).

O interesse pelos estudos sobre Alencar e sobre seu léxico fora reforçado pelo fato de que, como leitora, sentia a necessidade de conhecer o significado de muitas palavras que o autor utiliza. Por exemplo, quando lia trechos como

> O nome de Irapuã voa mais longe que o goaná do lago, quando sente a chuva além das serras;
> O búzio dos pescadores do Trairi e a trombeta dos caçadores do Soipé responderam.
> Todos os pescadores em suas jangadas seguiam o chefe e atroavam os ares com o canto de saudade, e os múrmuros do uraçá, que imita os soluços do vento. (ALENCAR, 1965).

Então perguntava-me "o que seriam **goaná**, **Trairi** e **uraçá**"? O desconhecimento destas palavras não me impedia de continuar a leitura, porém sentia a necessidade de conhecer os seus significados. Os dicionários de que dispunha não apresentavam definição da maioria das palavras utilizadas pelo autor e desconhecidas por mim, dentre os seus verbetes. Não encontrava meios para descobrir o que significavam.

Apesar de Alencar ter tido o cuidado de oferecer ao leitor um glossário com alguns vocábulos, cujo significado julgou importante esclarecer, isso ficou restrito a uma pequena

[2] Este trabalho é um artigo que deveria ser escrito com um tema diferente da tese. Trata-se de uma exigência para cumprimento de créditos para finalização do curso de Doutorado do Programa de Pós-Graduação em Estudos Linguísticos da Universidade Federal de Uberlândia.

parcela dos vocábulos, portanto eu ainda continuava com muitas dúvidas em relação a uma outra parcela de vocábulos utilizados pelo autor. Alencar também apresentou nas notas de rodapé ou no próprio glossário o processo de formação de alguns vocábulos. Sendo assim, é possível dizer que o autor se preocupava com a possível dificuldade de os leitores compreenderem seus textos, tanto no que se refere à sintaxe, como em relação à grafia das palavras, por isso optou por explicar alguns vocábulos em forma de notas.

Estudar o léxico de Alencar, por si só, não é um feito totalmente inédito, porém ao olhar as limitações dos trabalhos já existentes e analisados, permite-me uma nova proposta, cujos resultados, acredito, poderão contribuir com os estudos da linguagem, com o ensino do léxico e, mais especificamente, com os estudos lexicais da literatura brasileira.

Quantos olhares já perpassaram pelas obras de José de Alencar e quanto ainda há que se pesquisar e escrever sobre o autor? Por mais que possam parecer esgotáveis as análises e pesquisas sobre Alencar, há, ainda, novos olhares possíveis sobre sua obra. Esta pesquisa comporá o quadro de tantos outros estudos de pesquisadores que vão desde trabalhos de graduação a renomados críticos literários e a conceituados linguistas.

A reflexão saussuriana de que "é o ponto de vista que cria o objeto" (SAUSSURE, 1975, p. 15) referenda nosso intento, uma vez que José de Alencar pode ser considerado uma fonte inesgotável de pesquisa, dada a diversidade de sua produção. A obra literária de Alencar, por si só, já alavancou estudos sob diversas perspectivas e em várias áreas, assim como os textos não literários também já foram analisados sob perspectivas e reflexões diversas. A despeito desta diversidade, o olhar pelo qual optamos por empreender a nossa análise é um recorte inédito, baseando-se nos limites que minha investigação alcançou sobre trabalhos já desenvolvidos acerca do léxico de José de Alencar.

Especificamente, estudamos o léxico indianista de José de Alencar em suas obras também consideradas indianistas[3]: *O Guarani*, *Iracema* e *Ubirajara*. Em princípio, nosso intento era realizar uma pesquisa baseada em *corpus*, com o propósito de elaborar um Vocabulário com os vocábulos indianistas presentes nas obras mencionadas para disponibilização *online;* porém, com o decorrer do manuseio e observação dos dados, por meio do programa *WordSmith Tools* (doravante *WST*), outras possibilidades de estudo e de

[3] Nesta pesquisa não fazemos distinção entre as línguas indígenas, já que Alencar utiliza o Tupi, o Guarani e outras línguas indígenas na composição de seus vocábulos.

análises surgiram. Assim, com base nas afirmações de Berber Sardinha (2004; 2009) acerca das pesquisas guiadas por *corpus*, deixamos o *corpus* apresentar as possibilidades de análises.

Com base na indagação de que Alencar utilizaria um léxico específico em suas obras indianistas, se comparado às demais obras consideradas não indianistas, contrastamos, então, o léxico das três obras indianistas, que compõem nosso *corpus* de estudo, com nosso *corpus* de referência, composto pelas demais obras do autor. Com o fito de corroborar os resultados desse contraste, entre o *corpus* de estudo e o de referência do mesmo autor, utilizamos também outros três *corpora* de referência, que estão detalhados no Capítulo de Metodologia. Permanecemos com o intuito do Vocabulário indianista de Alencar para disponibilização *online*, entretanto, nesta tese, apresentamos uma proposta que é exemplificada com fichas lexicográficas e verbetes de 17 vocábulos.

A partir das orientações recebidas na banca de qualificação da tese, empreendemos nosso estudo também sobre a etimologia dos vocábulos indianistas de Alencar. Isto também porque, nas primeiras análises, percebemos que o autor criou palavras com base no seguinte procedimento: decomposição de alguns vocábulos e, com os fragmentos decompostos, composição de novos vocábulos, de acordo com suas pretensões literárias e linguísticas. Como se trata de étimos de Alencar, passamos a denominar o procedimento como Etimologia Ficcional Contextual[4]. O emprego da expressão Etimologia Ficcional Contextual se explica pelo fato de que o autor cria os étimos com o intuito de atender a um propósito de criação de palavras na história das obras, baseando-se no contexto da ficção de seus romances indianistas.

1.1 José de Alencar: alguns estudos

Como mencionado, há outros estudos sobre Alencar; contudo, diante da impossibilidade de se abarcar todos os trabalhos já realizados, escolhemos aqueles que se relacionam com nossa proposta de pesquisa. Para isso, estabelecemos um recorte dos estudos já desenvolvidos presidido pelos seguintes critérios: i) textos que abordem o léxico indianista, especificamente; ii) estudos sobre o léxico de Alencar, porém em perspectivas diversas da temática do indianismo; iii) aqueles que abordassem o indianismo em Alencar sem ênfase no

[4] A expressão Etimologia Ficcional Contextual será melhor explicada no Capítulo 3 do Referencial teórico.

léxico; por derradeiro, iv) aqueles que abordassem a metodologia e abordagem da Linguística de *Corpus* (LC) para análise do léxico do autor.

Em Ávila (2004), e como já apontado, elaboramos um glossário do léxico indianista de José de Alencar, tendo como referência a obra *Iracema*. Naquela ocasião, em razão da extensão do léxico, optamos por analisar apenas os nomes substantivos. O objetivo da pesquisa foi verificar se o léxico empregado por Alencar teria o caráter indianista sugerido. Para esse trabalho, utilizamos cinco dicionários para cotejar os vocábulos e organizar o glossário em ordem alfabética.

Por sua vez, Godoi (2006), apresentou uma proposta semelhante, em um artigo, que trata sobre o vocabulário indianista de *Ubirajara,* objetivando, com esse trabalho, evidenciar os traços ideológicos do autor ao criar um contexto metafórico e descritivo do ideário brasileiro. Segundo a autora, José de Alencar buscou no emprego lexical do indianismo um "abrasileiramento" vocabular e cultural. A autora fez um recorte de 14 palavras, consideradas representativas e as cotejou em cinco dicionários, a fim de verificar se, nesses dicionários, haveria a abordagem indianista e quais seriam as marcas de uso adotadas.

A partir da análise do texto de Godoi (2006), e considerando também a sua extensão, é possível realizar algumas observações. Inicialmente, a autora considerou representativas as 14 palavras selecionadas, porém não foi apresentada explicitação dos critérios de seleção e de informações sobre quantas palavras consideradas indianistas foram identificadas ao total. Outra questão a ser observada, no texto de Godoi (2006, p. 86), está relacionada à afirmação "Sabe-se que muitas dessas palavras são meras criações do autor e também, que muitas já se encontram dicionarizadas", uma vez que a autora não explicita a partir de qual estudo é feita esta afirmação.

Embora a temática que nos interessa seja o léxico indianista, trabalhos que versam sobre o léxico em Alencar em outras abordagens nos trazem à tona o quanto o autor foi alvo de estudos, assim como reforça a sua importância para a Língua Portuguesa falada no Brasil e para a literatura nacional. Por outro lado, como o foco desta pesquisa é também o léxico, torna-se procedente averiguar os estudos já realizados sobre o esse tema.

Queiroz (2006) pesquisou sobre o léxico alencariano de *O Sertanejo* numa perspectiva semântica. O objetivo do trabalho foi verificar aspectos do léxico utilizado por Alencar na

configuração do espaço geográfico do sertão caracterizado pelo autor. Para a análise do *corpus*, a autora utilizou o Sistema de Conceitos desenvolvido por Hallig e Wartburg (1963).

Semelhante à proposta de Ávila (2004), Queiroz (2006) analisou apenas os nomes substantivos da obra. Para isso, apresentou o vocábulo, a quantidade de ocorrências, uma passagem abonatória, as definições apresentadas pelos dicionários de exclusão e fez análise léxico-semântica de cada vocábulo selecionado. A autora conclui que Alencar tinha o propósito de atribuir à obra *O Sertanejo* um caráter nacional pela utilização de um léxico próprio do Brasil, especialmente com a ênfase em elementos da fauna e da flora. Concluiu, também, que o sertão, na obra alencariana, foi criado com o propósito de afirmar a identidade e a nacionalidade em contraposição aos lusitanos.

Com uma abordagem que vai ao encontro do nacionalismo apresentado por Queiroz (2006), Costa e Sales (2015) fazem um estudo sobre o léxico empregado por Alencar para descrever o vestuário feminino nas obras *Lucíola*, *Diva* e *Senhora*. As autoras fazem um recorte de 20 palavras para compor o glossário proposto no artigo. Segundo as autoras, o léxico utilizado por Alencar foi um recurso, por meio do qual ele expressa o nacionalismo. Para a seleção das unidades léxicas, as autoras se valeram do *software WST* na versão 5.0 e, como aportes teóricos, as orientações da Lexicologia.

Uma estudiosa de Alencar, que merece destaque, é a alemã Ingrid Schwamborn. A autora cursou Doutorado na Universidade de Bonn, Alemanha, e pesquisou sobre a obra alencariana. Posteriormente, publicou livros que foram traduzidos para o Português que tratam de todos os aspectos que envolvem a produção alencariana. O livro *O Guarani era um Tupi? Sobre os Romances Indianistas de José de Alencar,* Schwamborn (1998), publicado pela UFC – Universidade Federal do Ceará, pode ser considerado um tratado sobre Alencar de grande pujança, pois a autora propõe comparações entre os romances do autor e outras obras utilizadas, como fonte de pesquisa, pelo autor na época em que escreveu seus romances. Apresenta também fruto de pesquisas realizadas em bibliotecas brasileiras como na Biblioteca Nacional, onde pôde ter contato com os rascunhos de Alencar, para aprofundar suas análises.

Eckert e Röhrig (2016) apresentaram um trabalho em que analisam os antropônimos utilizados por Alencar em *Ubirajara*, a fim de verificar se existe relação entre o significado de alguns antropônimos utilizados por Alencar com o comportamento ou as características físicas desses personagens. Os autores buscaram fundamentos na Etimologia para as análises que possibilitaram concluir que existe relação entre o significado dos nomes e as características

dos personagens. Sendo assim, concluíram também que a escolha dos nomes analisados, por Alencar, não foi fortuita, mas etimológica e simbolicamente motivada.

Tendo em vista a pujança de textos sobre vida e obra de Alencar e diante da impossibilidade de inclusão de todo esse material, novamente recorremos à opção de um recorte, desta vez pelo indianismo, cujos trabalhos reportam ao indianismo de Alencar sem adentrar no universo do léxico.

O trabalho de Pagnan (2005) analisa aspectos da produção literária de José Alencar e Machado de Assis no viés da temática indianista, com o pressuposto de que a proposta dos autores de se construir a identidade do povo brasileiro foi frustrada. O autor concluiu que, apesar dos novos rumos para a tradição literária e língua brasileira, e em contradição à tradição europeia, o propósito de Alencar, por exemplo, não se concretizou. Isso porque não é mencionado o negro, outro representante brasileiro, apesar dos poucos vestígios da presença do negro como escravo, nas obras de teatro. Para Pagnan (2005), Alencar viu, apenas, no relacionamento entre índios e o português, a gênese do brasileiro, o que seria inverdade com a realidade.

Por outro viés, Pagnan (2005) destoa de muitas análises que enaltecem Alencar como um formador de uma linguagem genuinamente brasileira e como um autor que buscou fixar uma literatura independente da interferência europeia. Porém, e dada a engenhosidade da proposta, o texto se apresenta tímido em relação a esta discussão.

Com base nas buscas que realizamos, não identificamos estudos que analisassem o léxico indianista em José de Alencar pela perspectiva da LC. Portanto, podemos afirmar que não foram encontradas pesquisas anteriores à presente tese, envolvendo a metodologia e abordagem da LC (SCOTT, 2012), aplicados à análise do léxico indianista de Alencar. Além disso, tampouco identificamos qualquer trabalho que, pelo viés da LC, adotasse como critério de análise o contraste entre o *corpus* de estudo e o *corpus* de referência do mesmo autor.

Assim sendo, este trabalho levará em conta os elementos cuja combinação entendemos ser inovadora: i) léxico indianista de José de Alencar e Linguística de *Corpus*; ii) contraste do *corpus* de estudos com *corpus* de referência do mesmo autor; iii) vocabulário indianista de José de Alencar a ser disponibilizado em Vocabulário *online*, iv) análise etimológica do léxico indianista de Alencar. Podemos, com o exposto, antecipar a justificativa da escolha do

tema e da metodologia de pesquisa; contudo, na próxima seção, detalhamos a nossa justificativa desta pesquisa.

1.2 Justificativas da pesquisa e da tese

Para o desenvolvimento de uma pesquisa, há alguns fatores motivadores das decisões a serem tomadas. Assim sendo, apresento as justificativas para minha decisão de estudar o léxico indianista em José de Alencar.

Embora houvesse a preocupação de Alencar em tornar seus textos mais compreensíveis para os leitores por meio de notas, houve, antes, a escolha criteriosa do léxico para compor suas obras. Podemos ratificar em Hjelmslev (1966) citado por Beividas (2015), que o sentido é amorfo e toda a significação é decorrente e dependente do contexto situacional pleno, que somente existe com relação a ele. Em muitos casos, a prática lexicográfica tradicional desconsidera essa relação e apresenta os itens lexicais de maneira isolada, seguida dos significados.

Em outros casos, quando os lexicógrafos se preocupam com o contexto, apresentam frases que ilustram o significado em contextos abonatórios de obras diversas. Obras lexicográficas que levam em consideração um contexto em específico, como o indianismo em Alencar, não foram identificadas. Nosso intento não é desmerecer o valioso trabalho dos lexicógrafos, mas o de colaborar, apresentando um vocabulário dos vocábulos indianistas, considerando suas particularidades de emprego nas obras de um autor específico.

Outra justificativa para esta pesquisa é o fato de que José de Alencar é um ícone da literatura brasileira, tanto do período em que se inscreve, Romantismo, como na história da constituição literária do Brasil. Alencar busca inaugurar uma língua genuinamente brasileira, rompendo com os modelos impostos pelos europeus, isto é, busca romper com o eurocentrismo linguístico. Além disso, Alencar é um dos autores mais recomendados para estudos pelos currículos escolares, tanto do Ensino Fundamental quanto do Ensino Médio. Como também já foi sugerido para compor o grupo de obras literárias como conteúdo dos concursos de vestibulares de diversas universidades do país.

Acredito que estudar o léxico de uma língua ultrapassa as fronteiras do estudo das palavras dessa língua. O estudo do léxico de uma língua revela o seu próprio movimento, ou seja, a expansão, o acréscimo, a exclusão e a ressignificação das palavras que, por sua vez, revelam a visão de uma sociedade. Assim sendo, é no nível do léxico, que se podem evidenciar características de uma determinada sociedade no concernente ao espaço geográfico, à cultura e hábitos, enfim, é possível se traçar um perfil dessa sociedade. Esse perfil pode ser estabelecido considerando tanto o caráter sincrônico em que se pode estudar uma sociedade em um recorte do tempo, como também considerando o caráter diacrônico, que permite analisar as mudanças linguísticas, sociais e culturais, pelas quais passa uma sociedade no decorrer do tempo.

Estudar o léxico indianista na perspectiva da Etimologia, além de oferecer subsídios para a melhor compreensão das pretensões literárias do autor no contexto das obras indianistas, também proporciona melhor compreensão de itens lexicais que foram criados por Alencar e, atualmente, fazem parte da Língua Portuguesa.

Além do exposto, são raros os estudos sobre o léxico indianista de José de Alencar e aparentemente inexistentes aqueles que analisam o léxico indianista pelo viés da LC.

Ressaltamos que este estudo se justifica, também, pois análises dessa natureza podem ser úteis para as pessoas que têm interesse pelo estudo do léxico, especialmente o léxico indianista, como também para leitores de José de Alencar, especialistas que se interessam pelo assunto, assim como para os consulentes que necessitam de um vocabulário especializado do léxico indianista do autor. Também vale ressaltar que não foram encontradas, até o presente momento da escrita da tese, obras lexicográficas publicadas em meio digital, que versam especificamente sobre o léxico indianista de Alencar, como nos propomos nesta tese.

A partir das exigências impostas pelo objeto de estudo desta tese, a manipulação de inúmeros vocábulos do *corpus* de estudo e contrastá-los com outros *corpora* de referência, nosso estudo se realizou com o auxílio de ferramentas próprias da LC, conforme Berber Sardinha (2004; 2009), com procedimentos metodológicos específicos de compilação e análise de *corpora* textuais.

A escolha pelo programa *WST* se deveu à facilidade e agilidade na manipulação dos dados, por apresentar uma interface "amigável" para a análise de milhares de palavras e fornecer resultados robustos que podem ser considerados confiáveis. Os preceitos da LC

foram essenciais para a elaboração da proposta do vocabulário, pois possibilitaram a análise dos contextos abonatórios em busca dos étimos do autor, fornecendo subsídios para o processamento e tratamento dos *corpora* em contraste.

Para além das ferramentas do *WST*, no desenvolvimento da pesquisa, utilizamos como referência dicionários exclusivos de língua indígena e de língua portuguesa[5].

Ressaltamos que os dicionários de língua geral do português foram escolhidos em razão de datarem de um período anterior às publicações de José de Alencar.

1.3 Hipótese e questões de pesquisa

Referente aos aspectos relacionados ao léxico indianista de Alencar, a hipótese que deu início à presente pesquisa toma como base o princípio de que (1) Alencar utiliza um léxico específico em suas obras indianistas comparadas com as demais obras do autor consideradas não indianistas. A partir dessa hipótese, surge uma outra de que, considerando o léxico empregado em um contexto específico, que é o universo dos indígenas, (2) Alencar cria étimos a fim de atribuir aos seres e objetos nomeados não apenas um rótulo, mas também atribuir as características que estão impregnadas na etimologia do vocábulo criado.

Como mencionado, Alencar ambicionou um comportamento linguístico autenticamente brasileiro, sem o apego aos padrões do Português de Portugal; assim, procurou "criar" uma língua que retratasse a "fala brasileira". Mas esse desejo de uma língua brasileira não surgiu do acaso; Alencar percebeu que havia, no modo de falar do povo brasileiro, peculiaridades distintas em relação a Portugal. Com base nisso, o autor acreditou que a literatura também poderia retratar uma língua com características exclusivamente brasileiras e, nessa tentativa de valorização dessa língua, cabe perguntar, até que ponto Alencar não teria buscado transpor para o romance a "realidade" indígena?

Com base na indagação exposta, outros questionamentos que surgem são:

- Qual é a especificidade do léxico utilizado por José de Alencar em suas obras consideradas indianistas em relação a suas demais obras?

[5] Estes dicionários estão em versão PDF, disponibilizados pela Biblioteca Brasiliana da USP.

- Considerando a quantidade de notas explicativas do autor nos romances, Alencar cria étimos para atender a seus desejos literários e linguísticos?

1.4 Objetivos da pesquisa

Visando à análise dos étimos indígenas de Alencar e para responder às questões propostas, pretendemos:

1.4.1 Objetivo geral

- Realizar um estudo etimológico ficcional contextual sobre o léxico indianista de José de Alencar em O *Guarani*, *Iracema* e *Ubirajara*;

1.4.2 Objetivos específicos

Os objetivos específicos são:

i) Identificar e descrever o léxico específico de Alencar em suas obras indianistas;
ii) Realizar uma análise etimológica ficcional contextual com base nos itens lexicais indígenas de Alencar;
iii) Propor um desenho para elaboração de um Vocabulário com os vocábulos indianistas de Alencar, a ser disponibilizado *online*;

1.5 Organização da tese

Além desta Introdução e dos elementos antecedentes a ela, das Conclusões, das referências e dos apêndices, esta tese está organizada em 05 capítulos, a saber:

No capítulo 2, apresentamos o universo da pesquisa, por meio de uma breve biografia do autor José de Alencar, e traçamos considerações sobre o nacionalismo e o indianismo do autor e finalizamos o capítulo com apontamentos sobre a sua tríade indianista: O *Guarani*, *Iracema* e *Ubirajara*.

No capítulo 3, dedicamos nossa atenção aos postulados teóricos que sustentam esta pesquisa. Apresentamos, portanto, nosso entendimento em relação ao léxico; sobre a Lexicologia e Lexicografia; caracterização e constituição de obras lexicográficas e constituição de um dicionário. Continuamos os fundamentos teóricos com breves comentários sobre Campos semânticos e os pressupostos sobre etimologia, para finalizar com questões relacionadas à Linguística de *Corpus*.

O capítulo 4 destina-se a tratar do *Corpus* e da Metodologia de trabalho. Caracterizamos, assim, nosso *corpus* de estudo e o *corpus* de referências do autor José de Alencar e descrevemos os demais *corpora* de referência e o *Corpus* de consulta. Descrevemos de maneira pormenorizada os procedimentos metodológicos utilizados na pesquisa como levantamento dos *corpora*, armazenamento e preparação do *corpus*, extração das listas de palavras e análise contrastiva entre as listas, identificação do léxico indianista e agrupamento em campos semânticos.

A análise do *corpus* está dividida em três capítulos:

O Capítulo 5 trata da análise contrastiva do *corpus* de estudo com os *corpora* de referência, para identificação e extração do léxico indianista e consequente agrupamento em campos semânticos;

O Capítulo 6 é destinado à análise da Etimologia Ficcional Contextual do léxico indianista do José de Alencar;

Já o Capítulo 7 trata da proposta para elaboração de um vocabulário *online* com os vocábulos indianistas das obras também consideradas indianistas do autor. Nesse capítulo trazemos 17 fichas lexicográficas, acompanhadas de seus verbetes correspondentes e descrição da arquitetura do Vocabulário *online* proposto.

Por derradeiro, apresentamos as conclusões, as referências que sustentaram a teoria em que nos baseamos e, por fim, os apêndices.

O capítulo seguinte apresenta o universo da pesquisa e aborda as questões relacionadas ao autor José de Alencar como biografia, nacionalismo e indianismo e breve comentário sobre a tríade indianista: *O Guarani*, *Iracema* e *Ubirajara.*

CAPÍTULO 2 - UNIVERSO DA PESQUISA: ASPECTOS HISTÓRICOS

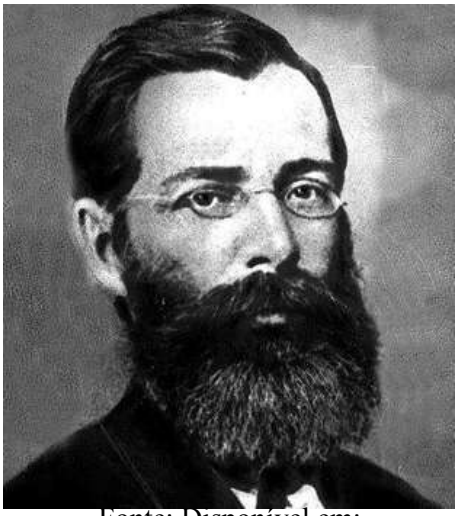
O tempo e o espaço se conjugam para compor um povo que habita um país ou, em uma instância micro, uma região, no que se refere às peculiaridades, assim como o que há em comum com outros povos. A paisagem, as manifestações religiosas, culturais e linguísticas, tudo é definido pela conjugação do tempo e do espaço em relação ao homem que vive em determinado momento. Neste capítulo, trataremos do homem e poeta José de Alencar e dos vários elementos que concorreram para que a sua vida e sua atuação não fossem esquecidas no cenário da literatura e da língua brasileiras. Exporemos, portanto, inicialmente, uma breve biografia do autor, para, em seguida, tratarmos do nacionalismo, seguido do indianismo, esse considerado um recorte da obra literária do autor. Finalizamos o capítulo, traçando comentários sobre a tríade indianista do autor, que compõe o objeto de estudo desta pesquisa.

2.1 José de Alencar: breve biografia, o nacionalismo e o indianismo

Esta seção está subdividida em três subseções que trazem alguns apontamentos sobre José de Alencar, no que ser refere à biografia, ao nacionalismo cultivado e ao indianismo defendido pelo autor. Apesar de viver poucos anos, a vida de Alencar é repleta de intensas e importantes realizações literárias e manifestações políticas, também manifestadas por meio da escrita, assim como pela vasta produção literária. Iniciamos, portanto, com os dados biográficos de José de Alencar.

2.1.1 José de Alencar: breve biografia

Imagem 1: José de Alencar



Fonte: Disponível em: https://www.estudopratico.com.br/biografia-e-obras-de-jose-de-alencar/>. Acesso em: 22 maio 2018

Em 1829, no primeiro de maio, nasce José de Alencar em Mecejana, no Ceará. Filho de Ana Josefina de Alencar e José Martiniano de Alencar. Pouco se fala de sua mãe, porém a avó Bárbara de Alencar é lembrada como heroína em razão de sua atuação política em prol do desligamento precoce de Pernambuco do poder colonial português durante a Revolução de 1817. Esteve, inclusive, na prisão, em razão de suas lutas pela liberdade. O pai de José de Alencar, José Martiniano de Alencar, foi uma figura vultosa na vida do autor pelo seu estilo patriarca dominante e pelo exercício em cargos políticos importantes. (SCHWAMBORN, 1998).

Em relação à literatura, a inspiração para a escrita se deveu ao fato de Alencar ter vivido em um ambiente familiar intelectualizado e favorável à formação cultural. Conforme o próprio Alencar relata em *Como e porque sou romancista*, lia, desde a infância, romances para as tias e a mãe. Consta também que Alencar viajava com os pais pelo interior do país e que as impressões dessas viagens refletiram e colaboraram para sua criação literária. Pode-se ratificar em Proença (1969, p. 43) que "o conhecimento do sertão deixará marcas inapagáveis na memória do menino José e será uma evocação permanente na obra do romancista J. de Alencar".

O seu pai fora transferido para o Rio de Janeiro para desenvolver atividades políticas e, evidentemente, a família o acompanhara. Alencar desenvolveu estudos no Colégio de Instrução Elementar. "É um aluno estudioso e que sabe ler muito bem; também em casa, é o leitor dos serões, em que a mãe, a tia e algumas pessoas mais se reúnem para ouvir romances de amor (...)". (PROENÇA, 1969, p. 44).

Estudou também em São Paulo, onde cursou francês, obteve aulas sobre Balzac, Victor Hugo e Chateaubriand, os quais seriam importantes aparatos literários para constituir suas obras, posteriormente. Em 1848, é transferido para a Faculdade de Direito de Olinda. Lá aprofundou seus estudos na biblioteca. Aproveitando também a proximidade com sua terra

Natal, volta ao Ceará, onde o reencontro com a paisagem reaviva-lhe o encantamento pelos índios e a admiração pela terra (PROENÇA, 1969).

Finalizado o curso de Direito, Alencar começou a trabalhar como advogado no Rio de Janeiro e, em 1854 e, a pedido de um colega de estudos, começa a escrever uma coluna semanal intitulada *Ao correr da pena*. A empreitada não foi somente realizações, porque a censura ao seu artigo sobre a nova febre de ações deixou-o aborrecido. Em 1855 assumiu o cargo de Redator Chefe do Diário do Rio de janeiro, então Alencar reassume a coluna *Ao correr da pena.*

A notoriedade de Alencar começou com a publicação de *Cartas sobre a Confederação dos Tamoios,* publicadas com o pseudônimo de Ig, no Diário do Rio de Janeiro, em 1856. Nas cartas, Alencar tecia ferrenhas críticas ao poema épico de Domingos Gonçalves de Magalhães, sobre o qual emitia críticas à qualidade e à pretensão da epopeia nacional brasileira. Alencar inicia uma crítica sistemática à obra no que se refere à frouxidão, às imperfeições métricas, à incapacidade de o autor dominar o tema e de realizar uma epopeia. Sobretudo, a crítica ocorreu pelo fato de Magalhães descrever os índios na perspectiva do homem civilizado.

A primeira experiência literária, em sentido próprio, surgiu, em fins de 1856, com o primeiro romance *Cinco Minutos*. Logo a seguir, em dez capítulos, Alencar escreve *A Viuvinha*. Porém, em 1857, iniciou a fase de popularidade do autor ao publicar, em folhetins, *O Guarani*, que se tornou o maior sucesso da história do jornalismo brasileiro. *O Guarani* foi publicado em 58 sequências, e, dia após dia, era esperado pelos leitores ávidos pela história de amor entre **Ceci**, a nobre moça portuguesa, e o índio **Peri**, da tribo dos Guaranis. Segundo Brito Broca (apud ALENCAR, 1965, p. xxiii), Alencar procurou demostrar como se devia fazer, em romance, a epopeia indianista do Brasil. O êxito prodigioso da obra diz da maneira feliz pela qual foi alcançado o objetivo.

Fora consagrado como um romancista. Depois volta-se a escrever peças para o teatro e o primeiro sucesso no teatro foi com a comédia *Demônio Familiar*. Esta peça foi seguida por *Verso e Reverso* e *Asas de um Anjo*. Esta última foi proibida pela polícia sob alegação de imoralidade. Alencar protestou no jornal e o caso repercutiu na Corte, no Rio de Janeiro, e até em São Paulo, de onde recebeu apoio de vários órgãos acadêmicos. Esse foi o primeiro dos

vários golpes em relação às questões literárias, pelos quais o autor passaria. Esses golpes lhe serviram de desalento literário.

A partir de *O Guarani,* outras obras surgiram para marcar, para posteridade, o talento de Alencar. As obras de Alencar e os anos de publicação estão detalhados no Capítulo 4 da tese *Corpus* e Metodologia. Proença (1969, p. 46) afirma que "dezembro de 1877, morre chorando, abraçado à esposa, preocupado com a pobreza em que vai deixar os seus". Apesar dos lamentos do autor e da morte prematura aos 48 anos de idade, deixa uma obra imensa, cuja repercussão somente, postumamente, foi percebida.

Há outros aspectos relevantes sobre a vida e a obra de José de Alencar, como também no que se refere à sua participação política, porém nos abstemos deste percurso, por não ser objeto desta pesquisa. Na próxima subseção, estão evidenciadas as ideias nacionalistas de Alencar

2.1.2 José de Alencar: o nacionalista

José de Alencar é considerado um dos grandes precursores dos ideais nacionalistas na prosa brasileira, por isso, também, é considerado um dos símbolos da literatura romântica no Brasil. Após a Independência do Brasil, 1822, período que coincide com o Romantismo brasileiro, os escritores tomaram consciência da necessidade de se construir uma literatura genuinamente brasileira, que se identificasse com as próprias cores e sabores, com as raízes históricas, linguísticas e, sobretudo, culturais.

Para se atingir o intento de constituir uma língua e literatura brasileiras sem influência do eurocentrismo, há que se iniciar, como já mencionado, pelo léxico e, na visão macro, a língua é o instrumento que melhor caracteriza um povo e que língua, sociedade e cultura estão intimamente entrelaçados. Seria então a língua o maior veículo para independência de um povo no que se refere à cultura, expandindo para a literatura e a política, desejos de José de Alencar. Nesse sentido, Câmara Júnior (1965) atesta que

> A língua é uma representação do universo cultural em que o homem se acha, e, como representa esse universo, as suas manifestações criam a comunicação entre os homens que vivem num mesmo ambiente cultural. A língua se apresenta como um microcosmo da cultura. Tudo o que esta possui se expressa através da língua; mas a língua em si mesma é um dado cultural. É fragmento da cultura de um grupo humano e sua língua. A língua é

> essencialmente a representação de um mundo extralinguístico em que os falantes se movem e que entra dentro de uma dada configuração formal. (CÂMARA JÚNIOR, 1965, p. 48).

Alencar é considerado um romancista distinto por ter transportado para suas obras a consciência de sua localização continental. Em suas obras, é possível reconhecer um esforço de resgate e fixação de uma identidade, o que o próprio autor explica no posfácio aos *Sonhos D'Ouro* que o povo brasileiro deveria reconhecer-se como nação e, para isso, valorizar e divulgar os costumes e a cultura até então submetidos à cultura do colonizador: "É lenta a gestação de um povo americano, que devia sair da estirpe lusa, para continuar no novo mundo, as gloriosas tradições de seu progenitor. Esse período colonial terminou com a independência". (ALENCAR, 1965, p. 101)

A nacionalidade e a fixação de uma identidade, aos olhos de Alencar, começariam pela literatura e, consequentemente, pela emancipação linguística. A língua é um dos principais instrumentos de dominação, portanto romper com os modelos linguísticos da literatura lusitana seria um indício de independência cultural. Segundo Preti (1977, p. 56), Alencar foi "o primeiro defensor da causa de uma 'língua brasileira' mais na prosa de ficção, no diálogo de suas personagens, mais de uma vez se manifestou sobre o purismo linguístico". Na busca por esse purismo, Alencar advogou sempre a "tese da existência no Brasil de uma língua nova, evoluída em relação aos padrões portugueses, por fatores extralinguísticos, língua que a literatura não poderia deixar de retratar". (PRETI, 1977, p. 56).

Por esse desejo, Alencar sofreu críticas ferrenhas de Pinheiro Chagas que percebeu defeitos nos romances do jovem autor. Pinheiro Chagas não foi condescendente com Alencar e com autores da época e via defeitos também em outros livros de escritores brasileiros. A crítica, principalmente, em relação à linguagem de Alencar, em decorrência, segundo Pinheiro Chagas, do emprego incorreto da Língua Portuguesa "ou antes a mania de tornar o brasileiro uma língua diferente do velho português, por meio de neologismos arrojados e injustificáveis, e de insubordinações gramaticais, que (tenham cautela) chegarão a ser risíveis se quiserem tomar as proporções duma insurreição em regra contra a tirania de Lobato". (PINHEIRO CHAGAS apud ALENCAR, 1965, p. 198).

Ciente de sua intenção e cônscio do que escrevia, Alencar não se retraiu, pelo contrário, defendeu-se das críticas de Pinheiro Chagas e de outros analistas críticos,

afirmando que “sempre direi que seria uma observação de todas as leis morais que a pujante civilização brasileira, com todos os elementos de força e grandeza, não aperfeiçoasse o instrumento de ideias, a língua (...)” e continuou se defendendo afirmando que “o Brasil possui uma língua sonora e abundante. A influência nacional já se faz sentir na pronúncia muito mais suave que nosso dialeto”. (ALENCAR, 1965, p. 171).

Outros autores como Franklin Távora (1965) e Antonio Henrique Leal (1965) também criticaram o estilo de Alencar. Távora (1965) considerou que seu estilo é inchado e as imagens se sucedem e se atropelam. Há também, segundo Távora (1965), um esbanjamento de imaginação e pelo fato de a linguagem não se aproximar da fala dos índios, por isso Alencar emprestou aos índios sua própria cultura. Leal (1965, p. 208) afirmou que “a sua linguagem e estilo são descuidados e por vezes desiguais e frouxo”.

Também Leal (1965)[6] acusou Alencar de estabelecer regras avessas ao uso geral da Língua Portuguesa, chegando a transgredir as normas da língua vigente, e defendeu a manutenção das lições dos “bons clássicos” portugueses, pois, para Leal, o Brasil é descendente de Portugal e, por isso, fala a mesma língua, assim

> é loucura tentar empresas tais, que só servem para o descrédito de quem o faz. Deixemo-nos de inovações extravagantes, onde já é miséria, e grande, não sabermos usar das riquezas que herdamos, para melhor recorrermos e admitir tudo o de que precisamos a fim de exprimir coisas ou novas, ou inteiramente brasileiras. (LEAL apud ALENCAR, 1965, p. 214).

Apesar das críticas ao estilo de Alencar, Leal (1965, p 208) considerou que, se compensados esses senões, estaríamos diante de uma obra com muitas belezas, tais como “a exatidão e firmeza de suas descrições, o bem sustentados dos diálogos, e as observações adequadas à feição verdadeiramente brasileira”. Por fim, afirma que “não carecíamos de mais ninguém para formar uma escola e pôr limites incontestes à nossa literatura”.

Uma outra análise, em relação ao estilo de Alencar, é proposta por Machado de Assis, que elogia o estilo, a linguagem e as imagens sugeridas, ao afirmar que

[6] Na edição de Iracema de 1965, edição do centenário, publicada pela Livraria José Olympio, o organizador compilou as críticas de autores sobre Alencar como Machado de Assis, Pinheiro Chagas, Franklin Távora, Henriques Leal. Compilou também outros estudos sobre Alencar e Iracema como os textos de Raquel de Queiroz, Agrippino Meyer, José Aderaldo Castelo e M. Cavalcanti Proença. Além dos textos considerados réplicas às críticas escritos por Alencar.

> O intuito era acertado; não conhecemos a língua indígena; não podemos afirmar se o autor pode realizar as suas promessas, no que respeita à linguagem da sociedade indiana, às suas ideias, às suas imagens; mas a verdade é que relemos atentamente o livro do Sr. José de Alencar, e o efeito que ele nos causa é exatamente o mesmo que o autor entende que se deve destinar ao poeta americano; tudo ali nos parece primitivo; a ingenuidade dos sentimento, o pitoresco da linguagem, tudo até a parte narrativa do livro, que nem parece obra de um poeta moderno, mas uma história de bardo indígena, contado aos irmãos, à porta da cabana, aos últimos raios do sol que se entristece. (MACHADO DE ASSIS apud ALENCAR, 1965, p. 189).

Reforçando os ideais de individualidade brasileira na língua e na literatura, Alencar buscou a construção e a difusão de um nacionalismo, sob alegação de que a língua falada no Brasil e em Portugal são diferentes. O autor utiliza uma comparação entre elementos da flora brasileira e de Portugal para se justificar que "um povo que chupa caju, a manga, o cambucá e a jabuticaba, pode falar uma língua com igual pronúncia e o mesmo espírito do povo que sorve o figo, a pera, o damasco e a nêspera?" Isso reforça a consciência do autor em relação aos seus romances, de tal forma que eles não são apenas um uso artístico da linguagem, mas uma criação artística com objetivos definidos de se criar uma individualidade, ainda que jovem, porém robusta, diferente das imposições lusitanas.

Uma das formas de se criar essa individualidade, para Alencar, foi valorizar o povo já habitante deste país, quando da chegada dos portugueses: o índio. Por esse motivo, o autor se propõe a escrever romances cujos personagens são também índios, assim como se propõe a defender o indianismo. Na subseção seguinte, abordaremos, brevemente, questões sobre o indianismo em Alencar.

2.1.3 José de Alencar e o indianismo

José de Alencar deixou uma vasta obra; contudo nos referiremos apenas às indianistas, por ser nosso objeto de pesquisa. Há vários estudos e inúmeras análises sobre as obras indianistas do autor, dentre elas, Castello (1979) propõe um ajuste cronológico nesses romances, classificando-os como grupo histórico-indianista:

> A aproximação das características dominantes, a histórica e a indianista, determina a subdivisão ou redistribuição das obras citadas em dois subgrupos: o primeiro, em que o traço indianista e mítico é predominante; o segundo, em que a perspectiva histórica é alimentada pelo sentimento de brasilidade que eclode no período da independência. Ressaltamos o primeiro subgrupo, em que se situa Iracema, assim constituído: Ubirajara, O Guarani e Iracema. (CASTELLO, 1979, p. 212).

Alencar, como estudioso da língua indígena, afirmou no prefácio de *Iracema* que "o conhecimento da língua indígena é o melhor critério para a nacionalidade da literatura. É nessa fonte que deve beber o poeta brasileiro; é dela que deve sair o verdadeiro poema nacional, tal como eu o imagino". O autor objetivou um texto verdadeiramente nacional e, em vários momentos, como em notas ou nos prefácios, justificou as características de suas obras no que se refere à forma e ao léxico.

Segundo Abreu (2011, p. 28), "a fixação desse tema (indianismo) sustenta-se em outros dois problemas fundamentais para a época: o nacionalismo e a instituição de uma literatura autenticamente brasileira". Pode ser considerado problema em razão das discordâncias exteriorizadas por diversos críticos que ora elogiam, ora depreciam o trabalho literário de Alencar. Ainda de acordo com Abreu (2011),

> o projeto de formação nacional da literatura ocorre por duas razões: 1) através dele o autor fundamentou as bases para a construção do seu indianismo e, portanto, discutiu, em confronto com outras visões, os problemas que lhe eram caros para tal finalidade; 2) por ser uma expressão consciente, depreende-se de tais debates a compreensão oitocentista acerca da arte literária, esta por sua vez articulada segundo um programa claramente delineado. (ABREU, 2011, p. 28).

Embora os critérios erigidos por Alencar para formação da nacionalidade tenham sido contestados por críticos e colegas literatos da época, os indicativos de nacionalidade e independência linguística são os mais concretos. Para isso, Alencar, além do emprego de vocábulos indianistas, também buscou traduzir, por meio da Língua Portuguesa, as características e ideias dos índios, como reforçou na carta à primeira edição do Livro *Iracema* que

> o poeta brasileiro tem de traduzir em sua língua as ideias, embora rudes e grosseiras, dos índios; mas nessa tradução está a grande dificuldade; é preciso que a língua civilizada se molde quanto possa à singeleza primitiva da língua bárbara; e não represente as imagens e pensamentos indígenas senão por termos e frases que ao leitor pareçam naturais na boca do

> selvagem. O conhecimento da língua indígena é o melhor critério para a nacionalidade da literatura. (ALENCAR, 1965, p. 141).

A tentativa de Alencar de reprodução das imagens e pensamentos indígenas não agradou a todos. Há também discordâncias entre autores ao analisarem o perfil dos índios nas obras de Alencar, por exemplo na perspectiva de Franklin Távora

> Sênio não compreende a poesia americana, como em geral tem sido concebida por bons talentos que o há precedido, e vem dar-nos o ideal da "*poesia verdadeiramente brasileira, haurida na língua dos selvagens"* na sua efeminada Iracema, onde os guerreiros falam uma linguagem débil, esmorecida e flácida, que não podiam de modo algum usar em sua braveza. (TÁVORA, 1872. p. 8-9. Grifos do autor).

Silviano (2006) também critica Alencar, porém reconhece a beleza e o trabalho cuidadoso do autor ao tentar uma nova escrita nacionalista. Segundo o autor:

> os diálogos entre os personagens são ao mesmo tempo castiços (eles se tuteiam), docemente falsos (as constantes comparações dos seus atos com o comportamento de diferentes animais e vegetais na selva), quando não são mera transcrição para o português de construção abonada pela sintaxe indígena. (SILVIANO, 2006, p. 56).

Diante da impossibilidade de citar todas as críticas a Alencar, cujas debilidades, no dizer de críticos, vão e vem, finalizamos esta seção com os comentários de Alfredo Bosi (2001), de que os personagens **Peri** e **Iracema** são passivos em relação à dominação do europeu, o que, segundo o autor, deveria ocorrer o contrário. O sentimento de servidão dos personagens "Nas histórias de **Peri** e de **Iracema** a entrega do índio ao branco é incondicional, faz-se de corpo e alma, implicando sacrifício e abandono de sua pertença à tribo de origem. Uma partida sem retorno". (BOSI, 2001, p. 178-179).

Encerramos esta seção sobre o homem José de Alencar para voltarmos o olhar para sua obra literária. Especificamente, trataremos da tríade indianista, que é, como já mencionado, objeto de estudo desta pesquisa. Isto, porque cônscio da vastidão da obra do autor, torna-se improcedente uma abordagem de toda a sua produção. Portanto, na seção seguinte, nosso foco se voltará para a tríade indianista de José de Alencar: *O Guarani, Iracema e Ubirajara.*

2.2 A tríade indianista de José de Alencar: *O Guarani, Iracema* e *Ubirajara*

José de Alencar, com o propósito de retratar o índio brasileiro, escreveu, como se sabe, a tríade indianista: *O Guarani* (1857), *Iracema* (1865) e *Ubirajara* (1874). A Imagem 02 mostra, na sequência, capa e folha de rosto da primeira edição de *Iracema*; capa da primeira edição de *O Guaran*i; e capa da terceira edição de *Ubirajara*.

Imagem 2: Capa e folho de rosto do livro *Iracema* (1ª edição) e capa de *O Guarani* (1ª edição) e capa da 3ª edição de *Ubirajara*

Fonte: Disponível em: Capa Iracema: http://g1.globo.com/pop-arte/noticia/2015/12/iracema-exemplar-da-1-edicao-do-livro-de-1865-vai-leilao-em-sp.html; Capa O Guarani: https://pt.wikipedia.org/wiki/O_Guarani; Capa Ubirajara: http://obuquineiro.com.br/produto/ubirajara-3a-edicao/.Acesso em 09 maio 2018.

Durante o seu percurso de vida, acompanhando sua família, Alencar, em visita ao Ceará, reencanta-se pelos índios e pelo conhecimento que eles possuíam da terra, íntimos das plantas e dos animais silvestres (PROENÇA, 1969). Esse encantamento foi o primeiro sinal para que, mais tarde, Alencar desejasse retratar os indígenas em sua produção literária. Apesar de muito tempo transcorrido entre a publicação do primeiro romance, *O Guarani*, e do último, *Ubirajara*, o conjunto remete à fundação, ainda que mítica, do Brasil.

Os três romances, embora o tema seja o mesmo, possuem características diferentes. Alencar mostrou seu forte poder de imaginação e pesquisa, transpondo para o terreno poético a "realidade indígena". O indianismo ocupou uma parte de sua produção literária e, portanto, é possível perceber uma evolução na criação das obras indianistas. O autor buscou estabelecer, em cada romance, um tipo de relação entre os índios e o povo não índio. Em *O Guarani*, o índio Peri convive no universo dos brancos; em *Iracema*, é o branco que vai habitar entre os índios; já em *Ubirajara*, o branco não está presente, isto é, índios convivendo

com os índios. Neste último romance, o índio é representado como "puro", sem interferência de culturas. Essa análise pode ser ratificada em Abreu (2011) ao afirmar que

> Ao compor um retrato idealizado do nativo, José de Alencar experimentou diferentes ângulos da temática indianista, já difundidos pela crítica do autor: em *O Guarani*, colocou o selvagem entre os portugueses; mais tarde, em *Iracema*, um português entre os selvagens; e, por fim, em *Ubirajara*, encontramos apenas índios entre os índios. (ABREU, 2011, p. 18).

Para Alencar, os índios seriam a melhor matéria para a epopeia nacional, assim sendo seria indispensável conhecer a língua indígena e reforça que é dessa fonte que deve beber o poeta brasileiro, pois ali encontraria o verdadeiro poema nacional. E reforça afirmando que o conhecimento da língua indígena é o melhor critério para a nacionalidade da literatura. "Ele nos dá não só o verdadeiro estilo, como as imagens poéticas do selvagem, os modos de seu pensamento, as tendências de seu espírito, e até as menores particularidades da sua vida". (ALENCAR, 1865, p. 141).

Assim é possível perceber que Alencar usou os elementos da natureza e dos índios para tecer as cenas de ficção. Abreu chama a atenção para

> a abertura dos enredos pelas chamadas "cenas da natureza": a virgindade e pureza da heroína; a função guerreira dos heróis; os obstáculos ao enlace amoroso, o emprego da cor local" que, com maestria, aproveitou para o enredo de suas obras, adequando-os aos seus propósitos de nacionalidade. (ABREU, 2011, p. 21).

A busca pela nacionalidade inspirou Alencar a narrar o índio, que, para ele, seria o melhor representante de um brasileirismo na linguagem e na literatura idealizadas por ele. Santiago (1984) propõe uma análise para *O Guarani,* afirmando que é o que tem "a situação histórica mais próxima da nossa", já sobre *Ubirajara*, o autor afirma que é "pré-cabralino". E afirma que mostra um Alencar

> com a determinação de se aprofundar mais e mais no conhecimento da cultura indígena, índio mesmo até o momento de sua pureza. À medida que passam os anos, sua visada se torna mais crítica e suas leituras dos cronistas do período colonial mais copiosas, enquanto o texto literário sai menos comprometido com os valores portugueses e mais engajado com as próprias descobertas nacionalistas. (SANTIAGO, 1984, p. 5-6).

Embora impregnado das tradições europeias e da tradição medieval, Alencar propunha uma linguagem e literatura brasileiras, Moraes Pinto (1995) analisa a tríade indianista afirmando que

> ele retoma apenas, no contexto de seus romances, os costumes medievais, evidentes n'O Guarani, um pouco menos em Ubirajara e bastante secundários em Iracema. E isso revela ainda uma vez a preocupação de fixar as origens de nossa formação, embora estabelecendo correlações com o romantismo europeu que buscava as raízes da nacionalidade na Idade Média. (MORAES PINTO, 1995, p. 38).

Ao encontro do mencionado por Moraes Pinto, Meyer (1979, p. 186-186) analisou uma passagem do romance na qual compara Peri com um cavaleiro, "crede-me, Álvaro, é um cavaleiro português no corpo de um selvagem! Peri pode ser considerado uma versão indígena de um cavaleiro sem mancha e sem medo". O índio Peri se submetia às mais perigosas situações para agradar a Ceci, como enfrentar uma onça e descer no precipício para apanhar um objeto perdido.

O Guarani é considerado a lenda de Tamandaré (o Noé indígena), pois a cena da enchente e quase dizimação se repete com Peri e Ceci, sobreviventes que, a partir de então, dariam origem a uma nova raça que povoaria o Brasil, à similitude da passagem bíblica de Noé.

Proença (1969) menciona que Taunay já divulgara a repercussão de *O Guarani*, publicado em folhetim em 1857, no Diário do Rio de Janeiro, o que despertou o entusiasmo da sociedade carioca. Taunay considerou esse romance uma verdadeira novidade emocional, até então desconhecida pela cidade. Menciona também que o romance despertou entusiasmo principalmente dos "círculos femininos da sociedade fina e no seio da mocidade". Enfatiza que o "Rio de Janeiro em peso lia *O Guarani* e seguia comovido e enlevados os amores de Ceci e Peri". (PROENÇA, 1969, p. 231).

Contudo, a repercussão do romance não foi privilégio entre os cariocas. Chegava também a São Paulo muitos dias depois, em razão do processo de transporte lento. Porém, ao chegar à capital paulista, os estudantes, embebidos de euforia e envolvimento, reuniam-se em repúblicas para ouvirem da voz alta de algum estudante assinante do *Diário do Rio de Janeiro* os capítulos da história do casal Peri e Ceci (PROENÇA, 1969).

Ainda de acordo com a análise desse autor, *O Guarani* teve maior importância pela prosa poética do volume, em que se reponta um novo estilo e um novo ritmo, o que transforma Alencar em um dos três renovadores da linguagem literária no Brasil, ao lado de Mário de Andrade e Guimarães Rosa.

O Guarani foi publicado primeiro e sob grande notoriedade, porém *Iracema* foi o romance que consagrou o autor. Em *Iracema*, não se pode conceber o indianismo simplificando-o a um gênero literário, mas trata de um novo estilo, sobretudo pela tentativa da fixação da imagem do Brasil. "*Iracema*, portanto, é tanto uma obra nitidamente universal, como o produto mais típico da nacionalidade e da autonomia de nossa literatura (RACHEL DE QUEIROZ, 1960, p. 60). Outros críticos reconhecem também a notoriedade de *Iracema* como Proença considera o romance um dos mais estudados pelos críticos e lido pelo povo.

> Com um século de existência, estudado por quase todos os grandes críticos brasileiros, negado por uns poucos, louvado e querido pelo povo, Iracema ainda será matéria de muitas pesquisas, estudo obrigatório de muitas gerações literárias. E será lido com enlevo pelo povo, que nele encontrará o lirismo, o amor e o sofrimento que as almas simples procuram na literatura. Esse é o destino das obras intrinsecamente ligadas ao sentimento dos povos. E *Iracema* é um desses livros. (PROENÇA, 1969, p. 53).

Apesar dos elogios advindos de alguns críticos sobre os romances de Alencar, houve também ferrenhas críticas de alguns puristas como Pinheiro Chagas, porém *Iracema* foi o mais bem-sucedido projeto de independência linguística e literária do Brasil em relação aos colonizadores portugueses. José de Alencar classifica *Iracema* como "lenda", a lenda do Ceará, por isso vislumbrou dramatizar a fundação da província do Ceará; transformando em símbolo da independência do país. Sobre isso, pode-se corroborar nas palavras de Castelo (1965) ao afirmar que

> Admitimos mesmo que é excepcional, dentro da Literatura Brasileira, mas explicável dentro do nosso Romantismo, o sentimento pátrio-local ampliado em proporções universais. Ao demais, quando o seu extravasamento é enriquecido pelas dimensões estéticas de uma criação de grande superioridade poética. É o caso da transubstanciação de um anseio afetivo no poema lírico que é Iracema, pelo autor dado como "Lenda do Ceará": muito mais lírico do que épica, os traços característicos que apresenta são a transbordante ternura, a saudade, o carinho, a abnegação e a renúncia, a dignidade, a altivez, a coragem, o desvelo heroico do homem e da mulher por um destino que se faz comum, o de dois povos que se unem, reduzidos ao instante inicial da formação da nacionalidade, entrevisto num espaço histórico afetivamente reduzido. Dessa maneira, a visão mais ampla das origens da própria nacionalidade deriva do sentimento íntimo local mais as

> impressões indeléveis do viver brasileiros, limitados este ao âmbito da experiência da sensibilidade aguda, debaixo de poderosa sugestão de luz, cor, som e relevo, para uma exaltação constante dos sentidos. E ele jamais esqueceria esse perene canto de verdadeira alegria, entoado pela Natureza, para sopitar as vicissitudes da sua vida. (CASTELO, 1965, p. 280).

A criatividade e o desejo de Alencar ao retratar o índio não cessou com *O Guarani* e *Iracema*. Completando a tríade indianista, o autor escreve *Ubirajara*, porém, como mencionado, não narra a convivência entre brancos e índios, retrata, portanto, o índio puro, representante das selvas brasileiras, o herói romântico. É o índio entre os índios. *Ubirajara* se diferencia dos demais, já que não há a presença do homem branco entre eles. É o índio idealizado e símbolo da literatura nascente, que volta às origens em busca de uma nacionalidade. "É seu orgulho de descender de um povo, aparentemente bárbaro, mas em verdade, apenas em estado de pureza e autenticidade humanas". (PROENÇA, 1969, p. 166).

O romance *Ubirajara* é narrado de acordo com os modelos da idealização romântica, porém valorizando a cultura brasileira por meio da representação da imagem do índio e de seus costumes. Alencar não se fixa no romance pelo romance, ele traz informações de cunho histórico e etnográfico, revelando o seu lado pesquisador dedicado à valorização do índio e da linguagem genuinamente brasileira (RAMOS, 2007).

Ainda há muito o que se falar sobre a tríade indianista de Alencar, tanto no que se refere ao estilo, como também ao enredo e vocabulário, porém encerramos neste ponto. A seguir, em consonância com os objetivos da tese, propomo-nos a trazer as questões teórico-linguísticas que embasam esta pesquisa.

CAPÍTULO 3 - FUNDAMENTOS TEÓRICOS

Neste capítulo, apresentamos o respaldo teórico-linguístico no qual se fundamenta nossa pesquisa. Para atingir tal intento, o capítulo foi dividido em seções, nas quais discorremos sobre Léxico, Lexicologia e Lexicografia, continuamos com seções destinadas à caracterização e à constituição de obras lexicográficas e a apontamentos sobre dicionários eletrônicos. A fim de alinhar com a abordagem e metodologia escolhida, fazemos breves apontamentos sobre Campos semânticos e dedicamos uma seção à Etimologia e à Etimologia Ficcional Contextual. Finalizamos o capítulo apresentando abordagens sobre a Linguística de *Corpus*.

3.1 Léxico

Estudar o léxico de uma língua é uma forma de conhecer a história social de uma determinada comunidade. Para Abbade (2009, p. 144), as palavras estão unidas em uma cadeia e "formam um campo linguístico através de um campo conceitual e exprimem uma visão do mundo de acordo com a reconstituição que elas possibilitam". Pode-se afirmar, então, que a língua revela a história de uma sociedade ou, neste caso, revela uma característica do autor José de Alencar, em um determinado conjunto de obras, as indianistas.

Estudar o léxico significa ir muito além do mero funcionamento da língua, significa adentrar na constituição e na organização de uma comunidade. O léxico é, então, um patrimônio cultural de uma língua e, como afirma Biderman (2001), é a soma de toda a experiência acumulada de uma sociedade e do acervo da sua cultura que perpassa por todas as idades. Assim, os membros dessa comunidade podem ser considerados "sujeitos-agentes" tanto no processo de reelaboração contínua do léxico, como no processo de perpetuação ou de exclusão desse léxico.

Ainda nesse sentido, Câmara Junior (1972, p. 266) afirma que a língua é um microcosmo da cultura. Para o autor, "tudo o que a cultura possui se expressa através da língua, mas também a língua em si mesma é um dado cultural, ao mesmo tempo em que a língua integra em si toda a cultura, ela deixa de ser esse fragmento para ascender à representação em miniatura de toda a cultura".

Por estar intimamente relacionado com a representação da realidade, todo o saber elaborado por uma comunidade, cristaliza-se por meio do léxico que está em constante movimento de renovação, ou seja, novas palavras são criadas ou ressignificadas, enquanto outras entram em desuso pelos falantes. Essa renovação ocorre porque o sistema linguístico é considerado aberto; sendo assim, os falantes são responsáveis pela criação de novas palavras ou pela ressignificação das já existentes, por meio da atribuição de novos significados. Biderman (2001, p. 179) ressalta que, no processo de desenvolvimento característico de qualquer sociedade, "o léxico se expande, se altera e, às vezes, se contrai. As mudanças sociais e culturais acarretam alterações nos usos vocabulares".

As alterações podem resultar no fato de que algumas unidades do léxico podem ser marginalizadas, ou seja, entrar em desuso ou desaparecer como, por exemplo, aquelas relacionadas a atividades que deixam de existir, como algumas profissões. Por outro lado, porém, coerente com o processo evolutivo da língua, outras unidades lexicais podem ser "ressuscitadas" com novas conotações. Ainda segundo Biderman (2001), o surgimento de novas significações de vocábulos já existentes ou a produção de novos é uma forma de enriquecer o léxico de uma língua.

Nesse sentido, pode-se ratificar em Antunes (2012) que

> A constante expansão do léxico da língua se efetua pela criação de novas palavras (doleiro, internetês), pela incorporação de palavras de outras línguas (deletar, mouse, leiaute, tuitar, blogar), pela atribuição de novos sentidos a palavras já existentes (salvar, fonte, vírus), processos que costumam coexistir e deixar o léxico em um ininterrupto movimento de renovação. (ANTUNES, 2012, p. 31).

Apesar da dinamicidade da língua, a renovação lexical não ocorre de maneira desordenada, ou seja, não se parte do zero, como afirma Antunes (2012, p. 29) "isso não quer dizer que as palavras sejam destituídas de toda e qualquer estabilidade de significado ou que, em cada momento de interação, os sentidos sejam criados inteiramente 'a partir de um estado cognitivo Zero'." Complementa o autor, afirmando que "às palavras são associados significados básicos, que constituem a base para a derivação de outros significados, próximos, associados, afins". (ANTUNES, 2012, p. 29).

É justamente por este caráter de instabilidade, de dinamicidade, de possibilidades múltiplas de criação e ressignificação é que a língua pode ser ajustada às necessidades dos

falantes sempre que necessário. Dessa forma, Biderman (1998, p. 91-92) afirma que "o léxico de uma língua constitui uma forma de registrar o conhecimento do universo. Ao darmos nomes aos referentes, o homem os classifica simultaneamente".

Os estudos sincrônicos de uma língua, por meio do léxico, revelam as características de um povo no que se refere aos hábitos e, de maneira ampla, revela a cultura desse povo, como já mencionado. Já os estudos diacrônicos do léxico de uma comunidade revelam as mudanças naturais pelas quais esta comunidade passou. Nesse sentido, Fiorin (2001, p. 115) afirma que "o léxico de uma língua se forma na História de um povo" e Vilela complementa afirmando que o léxico é "parte da língua que primeiramente configura a realidade extralinguística e arquiva o saber linguístico duma comunidade" (VILELA, 1995, p. 6).

Isso implica dizer que o léxico é a identidade do homem, pois é pelo léxico que o pensamento se manifesta, ou seja, a própria existência do homem está relacionada com léxico. O léxico pode ser definido então como um conjunto de palavras que compõem uma língua, ou seja, "designa o conjunto de unidades que formam a língua de uma comunidade, de uma atividade humana, de um locutor, etc." (DUBOIS, 2007, p. 364).

Os conceitos de alguns termos são relevantes nos estudos do léxico, portanto, ainda que sucintamente, são necessários esclarecimentos sobre os termos *palavra*, *vocabulário*, *lexema* e *lexia*. Dubois (2007) estabelece uma oposição entre léxico e vocabulário, ao afirmar que o léxico é reservado à língua e vocabulário ao discurso. Assim sendo, o vocabulário de um texto, por exemplo, pode ser considerado uma amostra do léxico de uma língua. Pode-se entender, então, que o léxico é o conjunto de unidades linguísticas básicas de uma língua e pode ser listado em ordem alfabética em dicionários, assim como o vocabulário também pode ser apresentado com as mesmas características de um dicionário, porém de uma amostra específica. Adotaremos, para esta tese, a perspectiva de Dubois (2007), de que vocabulário pode ser uma amostra específica com as mesmas características de um dicionário.

O vocabulário, no dizer de Pavel e Nolet (2002, p. 133), é o repertório monolíngue, bilíngue ou multilíngue de vocábulos organizados conforme critérios determinados, como palavras pertencentes a uma atividade específica ou de um determinado campo semântico, acompanhadas, normalmente, de definições com explicações sucintas.

A unidade-padrão de um vocabulário especializado é o *vocábulo*. Para Barbosa (2001, p. 40), vocábulo "tem um significado restrito e caracterizador de um universo de discurso (...),

assim o vocabulário deve armazenar elementos configurados de uma norma discursiva". Assim, procede falar em vocabulário de um autor, ou vocabulário em um grupo de obras pertencentes à mesma classificação como as indianistas de Alencar.

Quando se trata de definir *palavra*, Vilela (1995, p. 97) afirma que palavra possui um valor ambíguo, o "que torna desaconselhável o seu uso num discurso especializado. Em razão disso, prefere a utilização do termo unidade ou item lexical para nomear unidades consideradas semanticamente plenas e registradas no léxico de uma língua, como os verbos e substantivos. Essas unidades, segundo o autor, têm potencialidades de se combinarem entre elas ou com os afixos para formarem outras unidades e, nas relações entre elas, formar os sintagmas.

Apesar de o conceito de *palavra* ser impreciso, em razão de ter sentido amplo, alguns serão apresentados. Para Rey-Debove (1984), palavra pode ser formada por um só morfema, como mar; ou de vários, como declaração. Completando esse conceito, Azeredo (2004) explica que palavra é uma unidade geral a que corresponde o significado essencial. Vilela (1979, p. 19) amplia o conceito de palavra ao afirmar que a palavra "resulta da análise do sintagma mínimo, é uma entidade significativa solta, não extensa, embora frequentemente ainda analisável noutras entidades menores simultâneas". Ainda é possível analisar o vocábulo *palavra* de outras perspectivas como na análise morfológica ou fonológica, contudo, para fins desta tese, cabe-nos o conceito de que a *palavra* é uma entidade significativa.

A fim de evitar um possível problema de ambiguidade em relação ao termo palavra, Dubois (2007, p. 360), utiliza o termo *lexema* para definir a unidade de base do léxico, uma vez que, para o autor, "de um modo geral, o emprego do termo "lexema" permite evitar uma ambiguidade do termo 'palavra'. É embaraçoso ter que dizer que cantando é uma forma da palavra cantar, como o exige a gramática tradicional". O lexema é então o elemento da língua, ou seja, é a forma básica que possibilita as possíveis formas do discurso e todos os possíveis significados da palavra. Vilela (1979, p. 21) afirma que "o lexema é uma grandeza linguística real, de que dispõe a competência do falante/ouvinte, cujo alcance não é representável pelo uso, mas apenas pela reflexão".

Em relação ao termo *lexia*, Rey-Debove (1984, p. 48) afirma que "lexia é a unidade significativa máxima". O autor justifica esta afirmação dizendo que as lexias estão inseridas como unidades no código de nossa memória, porém não temos liberdade de mudá-las, por

isso as reproduzimos tais e quais. Pottier (1972) conceitua lexia como a unidade de comportamento léxico e é a unidade funcional significativa do discurso. O autor propõe que as lexias são classificadas em lexias simples, compostas, complexas e textuais. As lexias simples podem ser uma palavra como *mesa, cama;* já as lexias compostas podem ter duas ou várias palavras e são resultados de uma integração semântica, como *guarda-roupa, pé-de-moleque*; as lexias complexas são sequências de unidades consideradas como sequências cristalizadas pela língua e grafadas como uma unidade e correspondem a um único referente no plano da língua, como *guerra de nervos*; já as lexias textuais ocorrem quando uma lexia complexa alcança o nível de um enunciado de texto, como os provérbios ou adivinhações, por exemplo.

Sobre a classificação das *lexias*, Borba (2003, p.22, 23-24) também aponta que há lexias simples e complexas. Assim, as lexias simples são "formadas por uma única forma livre [cara, porto, vento] e complexas as que combinam mais de uma forma livre [porta-luvas, mal-mequer, joão de barro] ou uma forma livre e uma ou mais de uma forma presa [desconsolo, incontrolável]". O autor ainda complementa com a definição para palavras compostas que, segundo ele,

> Para as palavras compostas ou simplesmente compostos, têm sido propostas análises ao nível da morfologia derivacional. São itens lexicais complexos formados por justaposição de formas livres, cuja integridade fonética permite que sejam grafados com ou sem hífen, com ou sem espaço em branco: pé-de-cabra, bem-te-vi, sempre-viva, casa de saúde, casa da sogra, girassol, passatempo, varapau. (BORBA, 2003, p. 23-24).

Como já mencionado, a conceituação de vocábulos e suas especificidades é um procedimento conflituoso, porém estabelecer o recorte utilizado é essencial para situar o aporte teórico no qual se embasa esta pesquisa. Ressaltamos que, neste estudo, em razão de tratarmos acerca do léxico de Alencar, utilizamos *unidade lexical* ou *item lexical* na perspectiva de Vilela e *vocábulo,* na perspectiva de Barbosa, como sinônimos. A seguir, traçaremos apontamentos sobre Lexicologia e Lexicografia.

3.2 Lexicologia e Lexicografia

O léxico pode ser compreendido como uma ampla área de investigação das Ciências da Linguagem. Por se tratar de um sistema aberto, há uma dificuldade de estudá-lo em sua plenitude. A despeito dessa dificuldade, há estudos diversos sobre o léxico de um sem número de línguas. A Lexicologia, a Lexicografia, a Terminologia, a Terminografia e a Onomástica são as chamadas Ciências do léxico. Ater-nos-emos, apenas, por questões de pertinência com esta pesquisa, à Lexicologia e à Lexicografia, que designam duas atitudes e dois métodos em face dos estudos do léxico. A Lexicologia trata do estudo científico do léxico, já a Lexicografia é a ciência dos dicionários. Apesar de tratarem do mesmo objeto, o léxico, "ambas possuem um objeto de estudo, metodologia e pressupostos teóricos distintos, embasados na peculiaridade com que cada uma submete a palavra". (ÁVILA; MARTINS, 2008, p. 51).

Uma definição possível para Lexicografia, encontra-se em Dubois (2007, p. 367) ao afirmar que "Lexicografia é a técnica de confecção dos dicionários e a análise linguística dessa técnica". Lexicografia, porém, é um termo que apresenta mais de um significado, por um lado pode ser compreendido como os estudos dos significados das palavras e, por outro, o trabalho de compilação de dicionários. Dubois (2007, p. 367) afirma que "o termo é ambíguo, como lexicógrafo, que pode designar, ao mesmo tempo, o linguista que estuda a lexicografia e o redator de um dicionário". O autor explicita que é possível distinguir "a ciência da lexicografia e a prática lexicográfica e, do mesmo modo, o linguista lexicógrafo e o autor de dicionários". Ainda sobre o aspecto ambíguo do termo Lexicografia, Borba afirma, na mesma perspectiva de Dubois, que

> A lexicografia pode ser vista sob duplo aspecto: (i) como técnica de montagem de dicionários, ocupa-se de critérios para seleção de nomenclaturas ou conjuntos de entradas, de sistemas definitórios, de estruturas de verbetes de critérios para remissões, para registro de variantes etc.; (ii) como teoria, procura estabelecer um conjunto de princípios que permitam descrever o léxico (total ou parcial) de uma língua, desenvolvendo uma metalinguagem para manipular e apresentar informações pertinentes. (BORBA, 2003, p. 15).

Dessa forma, a Lexicografia ainda denomina o estudo dos dicionários sob várias perspectivas, como os princípios de produção de um dicionário bilíngue ou sobre as regras de organização de uma lexia, por exemplo. Por outro lado, denomina também os procedimentos

de produção de um dicionário, o qual pode ser considerado a obra maior da lexicografia. Da mesma forma, o lexicógrafo pode ser denominado tanto aquele que estuda os dicionários quanto aquele que escreve um dicionário. Não significa que ambos estejam apartados por seus trabalhos, mas é possível que as pesquisas realizadas por um estudioso sirvam de material para aquele que confecciona o dicionário. Villalva e Silvestre (2014, p. 187) atestam que "se considerarmos a compilação de dicionários fundados em trabalho lexicológico cientificamente validado, a lexicografia é um resultado da lexicologia aplicada".

Sobre a função da Lexicologia, Vilela (1995) afirma que ela serve para ajudar os leitores a interpretar textos, em primeira instância; e ajudar os consulentes a produzir textos, em segundo plano. Ressalta-se que a leitura dos itens lexicais de um dicionário, por si só, não garante a interpretação ou a escrita de um texto, mas auxiliam na definição de um determinado uso, na descrição ou no conceito desse item lexical. Para a materialização do traço sêmico, é necessário que o item lexical esteja inserido em um contexto, já que os textos não são formados por palavra isoladas.

No que diz respeito à definição dos itens lexicais, Villalva e Silvestre (2014, p. 12) afirmam que a palavra no léxico "é uma entidade abstrata e que reúne todas as informações que todos os domínios da gramática consideram relevantes". Daí a justificativa da necessidade de os dicionários estarem em constante alargamento, a fim de melhorar a qualidade das anotações. Os autores advertem que, apesar das falhas e lacunas, os *corpora* lexicais existentes são instrumentos de consulta obrigatória, sempre que se pretenda ou necessite verificar o uso de uma determinada palavra ou qual o significado com que essa palavra é usada.

Se à Lexicografia cabe o fazer e o estudo dos dicionários, à Lexicologia cabe o estudo do léxico e privilegia a palavra, como seu objeto principal de estudo. Sendo assim, a Lexicologia estuda as palavras em todos os seus aspectos, por se tratar de um estudo científico do léxico. Pode-se atestar em Barbosa (1991) que a Lexicologia se propõe a estudar todas as palavras de uma língua, nos mais diversos aspectos como verificar sua estruturação, seu funcionamento e sua mudança. Caberia, então, entre outras tarefas, definir conjuntos e subconjuntos lexicais; verificar a relação do léxico de uma língua como o universo natural, social e cultural; conceituar e delimitar a unidade lexical de base – lexia – bem como elaborar os modelos teóricos subjacentes às suas diferentes denominações; abordar a palavra como instrumento de construção e detenção de uma "visão de mundo", de uma ideologia, de um

sistema de valores, como geradora e reflexo de sistemas culturais; e analisar e descrever as relações entre expressão e o conteúdo das palavras e os fenômenos daí decorrentes.

A Lexicologia, por se ocupar do léxico das línguas de maneira integrada e completa, permite caminhos diversos como estudos relacionados à Fonética, à Fonologia, à Semântica, à Morfologia, à Sintaxe e à Pragmática. Portanto, a importância da Lexicologia está fundamentada nas investigações das estruturas do léxico, a partir da hipótese de que o léxico de uma língua se organiza a partir de leis estruturais. Sendo assim, trata-se de uma ciência de múltiplas investigações, inclusive sensível a fatores sociais e culturais.

Após os breves apontamentos sobre léxico, Lexicologia e Lexicografia, cumpre-nos tratar acerca da especificidade e tratar da constituição de dicionários, o que abordaremos na seção seguinte.

3.3 Caracterização de obras lexicográficas

A Lexicografia, como mencionado, pode ser o estudo e o fazer dos dicionários. Apesar da aparência de catálogo lexical, pela forma de listagens em que as unidades lexicais de uma língua são registradas, o dicionário possui regras próprias de confecção. São textos metalinguísticos e assumem uma função didático-pedagógica, como afirma Dubois (1971), o objetivo dos dicionários é essencialmente pedagógico, pela propriedade e possibilidade de levar conhecimentos aos leitores sobre uma comunidade inteira, por meio de sua língua.

O dicionário de uma língua pretende propor uma descrição do léxico de uma determinada língua, assim, como atesta Vilela

> A comunidade linguística codifica o seu saber acerca do mundo no seu dicionário de um modo prático, mas não reflete de modo total a língua, nem a própria realidade armazenada. Aliás, há outros testemunhos acerca do "arranjo" que a língua representa: a comunidade segmenta linguisticamente a realidade de acordo com os seus interesses religiosos, culturais econômicos etc. (VILELA,1995, p. 95).

O dicionário de uma língua, então, deve estar de acordo com as necessidades do usuário. Em razão disso, Vilela (1995) propõe que a definição lexicográfica deverá ser regida por algumas propriedades: i) a definição deverá estabelecer uma relação entre o geral

(gênero) e o individual (espécie); ii) a definição não poderá ser formulada negativamente, se for possível formulá-la positivamente; iii) a definição não poderá ser circular; iv) deverá apresentar uma correspondência gramatical entre a entrada e a paráfrase de definição; v) os elementos linguísticos usados na definição deverão ser mais frequentes do que o vocábulo definido.

Essas são regras gerais para elaboração de um dicionário, porém Villalva e Silvestre (2014) ressaltam que a descrição lexicográfica vai refletir os propósitos do lexicógrafo, como os métodos de compilação e os documentos disponíveis, ou seja, é um recorte condicionado pela perspectiva do autor. Os autores, por outro lado, reforçam que nenhum dicionário moderno ou antigo pode ser considerado como uma representação extensa e exata de um léxico, eles apresentam uma perspectiva subjetiva da língua e do léxico, num determinado período histórico, pois é possível afirmar que, a despeito do trabalho exaustivo, dificilmente se conseguirá reunir todas as palavras em uso de uma língua em um dicionário.

Nessa perspectiva, Bidermam (2001, p. 97) reforça a ideia de que "grandes monumentos lexicográficos de muitas línguas nada mais são que vastos repertórios vocabulares de um determinado estado da língua, pois um sistema aberto em expansão, como o léxico, não pode ser apreendido, nem descrito na totalidade". O dicionário é, então, o conjunto de vocábulos de uma determinada língua ou, em outra instância, pode ser de termos próprios de uma ciência ou arte, que mantém as características de os termos virem dispostos em ordem alfabética e com o significado ou a descrição do termo (MACIEL, 2001, p. 42).

Como os dicionários podem ser elaborados com funções específicas, Maciel (2001) ressalta que há tipologias para os dicionários e apresenta três: dicionário de língua geral, dicionário especializado e dicionários técnicos.

O dicionário de língua geral ocupa-se em abarcar o léxico de uma língua, portanto tende a apresentar um inventário mais completo. Segundo o autor, esse dicionário apresenta uma realização mais ou menos padronizada da unidade lexical e procura abranger os possíveis usos e significados. Pode incluir as formas que as palavras podem assumir, assim como paradigmas morfológicos, campos semânticos, sinônimos e antônimos, dentre outros aspectos. Com as informações constantes desse dicionário, espera-se que o usuário/leitor possa ser capaz de empregar as palavras no sistema linguístico, utilizando-a eficientemente para exteriorizar os pensamentos no processo de comunicação.

Já o dicionário especializado centra-se em um recorte de um léxico da língua geral ou da linguagem de especialidade, e foca em vocábulos especiais usados em contextos particulares. Apesar de ser um recorte da língua geral, o dicionário especializado mantém as regras do dicionário geral para elaboração e produção. O que os distingue é a questão de limites do objeto, já que o dicionário especializado faz um recorte e seleciona um conjunto de palavras pelos mais diferentes critérios, por representar um repertório linguísticos utilizado por um autor, por um grupo sociodialetal, ou, ainda, por grupos de especialidades.

Por último, na proposição de Maciel (2001, p. 44), há os dicionários técnicos que podem, por sua vez, serem classificados em dois grandes grupos: dicionários lexicográficos e dicionários terminológicos. Nos lexicográficos, "o termo é um elemento linguístico de um vocabulário especializado que tem seu referente no universo real ou imaginário. A definição do dicionário técnico lexicográfico é uma explicação ou descrição do significado de um termo como um item lexical dentro do sistema linguístico". Nos dicionários terminológicos, "o termo é um sistema de conceitos específicos". A definição parte de um corpo de conhecimento e inclui expressões complexas na forma de sintagmas, siglas, abreviaturas, ou seja, elementos que reportem às ciências envolvidas ou nos ramos que se desejam inventariar no dicionário.

Os dicionários gerais de língua estão em constante alargamento em busca de abarcar os itens lexicais existentes, assim como para melhorar as informações e anotações sobre os itens. Apesar disso, Villalva e Silvestre (2014) alertam que os dicionários são instrumentos de consulta obrigatória, quando se deseja averiguar o uso e o significado de uma palavra. Os autores também advertem que os dicionários de língua contemporânea são importantes instrumentos de consulta e de trabalho, porém há que se proceder a uma leitura crítica dos dados oferecidos, isso porque, em alguns casos, apresentam repetições de informações de dicionários antigos.

As advertências dos autores são pertinentes, porém hoje os dicionários ainda são uma rica fonte de atestação das palavras. Mesmo com as repetições das informações de dicionários antigos, os contemporâneos apresentam um esforço de adaptação do vocábulo à atualidade, já que, em alguns casos, a grafia, os contextos de produção podem ser diferentes, assim como a palavra pode ter sido ressignificada, o que exigiria uma readaptação.

Em relação aos dicionários, os consulentes tendem a se comportar de formas diferentes, ou seja, procuram informações para suprir a necessidade de um momento, por isso, para a produção de um dicionário, há critérios pré-estabelecidos como o público-alvo a que se destina. Em razão disso, há os dicionários monolíngues, bilíngues ou multilíngues. "Os dicionários monolíngues são tipicamente concebidos para utilizadores nativos da língua em objeto. Nos dicionários bilíngues, distinguem-se uma língua de partida e uma língua de chegada". (VILLALVA, SILVESTRE, 2014, p. 194).

Em relação aos dicionários monolíngues, Biderman (2001) atesta que

> Existem vários tipos de dicionários monolíngues, os dicionários de língua, os dicionários analógicos (ou ideológicos), os dicionários temáticos ou especializados (de verbos e/ou regência verbas, de sinônimos e antônimos) os dicionários etimológicos, os dicionários históricos, os dicionários terminológicos das diferentes áreas do conhecimento. (BIDERMAN, 2001, p. 131).

Por meio da afirmação da autora, é reforçado que os dicionários monolíngues são indispensáveis para consulentes que desejam esclarecimentos sobre itens de sua própria língua. Por outro lado, Dubois (1971) afirma que, para a consulta a outras línguas, parte-se para consultas a dicionários bilíngues ou plurilíngues, o que implica conhecimento pelo leitor seja de uma linguagem-fonte (tema) seja de uma linguagem-alvo (versão).

Em outro viés de análise sobre os dicionários, a apreciação quantitativa, os dicionários podem ser classificados com base no número de entradas, por exemplo, o Houaiss apresenta 442 mil verbetes, enquanto o mini Aurélio, cerca de 30 mil. Dubois (1971) afirma que os dicionários podem ser considerados extensivos, quando tendem a cobrir a totalidade dos itens lexicais de uma língua, de um dialeto ou léxico funcional de uma atividade técnica determinada; em contrapartida, há os considerados intensivos, quando se procede a uma escolha definida de itens. O objetivo é apresentar o maior número possível de informações, definições, desenvolvimento descritivo, histórico, entre outros.

Além do dicionário, como obras lexicográficas, há também o vocabulário e o glossário. Barbosa (2001) propõe a distinção entre *vocabulário* e *glossário* por critério quali-quantitativo. O vocabulário representa o universo do discurso; o glossário pretende representar o léxico de um único texto em sua especificidade A autora considera também que os dicionários estão no nível do sistema (língua, abstrata), todo o léxico é fonte do trabalho e

se tenta apresentar todas as acepções de um verbete, é o que se deseja; já os vocabulários, por outro lado, estão no nível da norma (fala do grupo, abstrata). O vocábulo é um recorte do sistema e o termo é uma unidade de trabalho relacionada a uma área de especialidade. Quanto às acepções, são todas mencionadas, porém considerando uma área de especialidade; e nos glossários, no nível da fala (individual, concreta), as palavras com significado específico são o foco do trabalho e, diferentemente do vocabulário, apresentam uma única acepção.

Fromm (2004), com base em Barbosa (2001), propõe um esquema acerca das tipologias, conforme Quadro 1.

Quadro 1: Tipologias de obras lexicográficas com base em Barbosa

Dicionário	**Vocabulário**	**Glossário**
Nível do sistema	*Nível da norma*	*Nível da fala*
Trabalha com todo o léxico disponível e o léxico virtual	Trabalha com conjuntos manifestados dentro de uma área de especialidade	Trabalha com conjuntos manifestados em um determinado texto
Unidade: lexema (significado abrangente; frequência regular)	Unidade: vocábulos/termos (significado restrito; alta frequência)	Unidade: palavras (significado específico; única aparição)
Apresenta (teoricamente) todas as acepções de um mesmo verbete	Apresenta todas as acepções de um verbete dentro de uma área de especialidade	Apresenta uma única acepção do verbete (dentro de um contexto determinado)
Perspectivas: diacrônica, diatópica, diafásica e diastrática	Perspectivas: sincrônica e sinfásica	Perspectivas: sincrônica, sintópica, sinstrática e sinfásica

Fonte: Barbosa, (2001, apud FROMM, 2004, p. 17)

Ainda com o propósito de estabelecer a distinção entre vocabulário e glossário, trazemos Krieger e Finatto (2004, p. 51), que afirmam que "glossário costuma ser definido como repertório de unidades lexicais de uma especialidade com suas respectivas definições ou outras especificações sobre seus sentidos. É composto sem pretensão de exaustividade". Pode-se afirmar, então, que glossário pode ser considerado um repertório, a partir de uma determinada especificidade como, por exemplo, glossário de uma obra literária ou de um autor.

Com base nas questões sobre tipologias de obras lexicográficas, consideramos que o produto de nossa pesquisa, constitui-se como um vocabulário, uma vez que tratará do léxico indianista que reflete a produção alencariana, no que se refere aos romances da trilogia indianista. Outro aspecto relevante sobre as obras lexicográficas é a sua constituição. Esse assunto é tratado na seção seguinte.

3.4 Constituição de obras lexicográficas

O trabalho de um lexicógrafo, na produção de uma obra lexicográfica, requer uma boa estrutura. Há de se criar um método para se extrair as fontes, para se redigir o texto no que se refere aos verbetes, ao contexto abonatórios, à etimologia, enfim todos os dados a serem disponibilizados. Além disso, o lexicógrafo deve ser um conhecedor da língua com a qual se deseja trabalhar e, no caso dos dicionários bilíngues ou multilíngues, deve ter conhecimento de ambas as línguas. Dentre os conhecimentos necessários que o lexicógrafo deverá possuir, destacam-se a pronúncia, a estruturação interna da língua, os processos de formação, o vocabulário, a sintaxe e a estilística, além dos conhecimentos técnicos sobre informática (SCHIERHOLZ, 2012).

Para se elaborar uma obra lexicográfica, segundo Schierholz[7] (2012), há de seguir algumas etapas. Inicialmente, o planejamento e a concepção de um projeto, para, ao mesmo tempo, criar uma infraestrutura para o desenvolvimento do conteúdo. Juntamente ao planejamento, determinar a abrangência do dicionário que deverá levar em conta a seleção dos itens lexicais, quantidade de itens e de informações a serem disponibilizadas e, por fim, organizar a forma de divulgar o produto.

Ainda sobre a constituição de obras lexicográficas, Cunha (1998, p. 15) afirma que, para se elaborar um dicionário de Língua Portuguesa, com base em princípios históricos, requer o "preparo de um corpo de lexicógrafos devidamente treinados nas modernas técnicas lexicográficas" e complementa afirmando que se acrescenta a isso "a execução de um programa, tanto quanto possível minucioso, em que se estabeleçam os critérios que devem

[7] Tomaremos o que disse o Schuerhikz (2012) e Cunha (1998) sobre a constituição de dicionários como sinônimo de constituição de obras lexicográficas, em razão de os procedimentos serem os mesmos, como também pelo fato de tratarmos, nesta tese, da elaboração de um vocabulário.

nortear o levantamento do vocabulário dos textos de língua portuguesa". Portanto, nos procedimentos de elaboração de uma obra lexicográfica, o lexicógrafo deverá recorrer aos modernos recursos tecnológicos.

Outra questão apontada por Schierholz (2012), sobre a constituição de obras lexicográficas, diz respeito ao processo de revisão, no qual se distingue a simples conversão para um meio eletrônico para um banco de dados para que o usuário possa pesquisar. Isso requer uma análise das possibilidades de busca que o consulente poderá escolher. Normalmente, o lexicógrafo conta com textos eletrônicos, impressos ou orais e até outros dicionários para organização de sua obra. O banco de dados é importante para se averiguar a frequência das palavras e, com isso, elencar aquelas que comporão o produto. Isso porque quanto maior a ocorrência de uma determinada palavra, maior a probabilidade de busca por ela. O contrário também ocorre, ou seja, quanto menor a ocorrência de uma palavra no banco de dados, menor a chance de os consulentes procurarem por ela.

Mesmo diante de todos os aparatos tecnológicos e do banco de dados do qual se selecionam as palavras, há, ainda, a questão da dimensão do dicionário, ou seja, quantos verbetes deverão ser inseridos na elaboração do produto. Essa questão faz parte dos propósitos do lexicógrafo ao estabelecer o planejamento para seu trabalho. De acordo com a quantidade de verbetes inseridos em um dicionário, ele é classificado de uma maneira. Biderman (2001, p. 130) propõe uma classificação para os dicionários, baseando-se no tamanho:

- Dicionário infantil e/ou básico – para crianças de até 7 anos devem conter aproximadamente 5.000 verbetes; para crianças entre 7 e 10 anos, os dicionários devem conter 10.000 verbetes;

- Dicionário escolar e/ou médio – para pessoas acima de 10 anos, entre 10.000 a 30.000 verbetes;

- *Thesaurus*: são considerados os dicionários mais robustos e devem conter entre 100.000 a 500.000 verbetes;

Ressalta-se, porém, que, mesmo que o dicionário contenha 500.000 verbetes, ele não conseguirá abarcar todo o léxico de uma língua. Como o exposto, o lexicógrafo terá que realizar, com base em princípios, a seleção das palavras que comporão os verbetes. Além

disso, deverá levar em consideração aspectos como a macro e microestrutura, sobre as quais serão traçados alguns apontamentos, a seguir.

3.4.1 Macroestrutura

A definição e explicitação de macroestrutura ainda é alvo de algumas controvérsias. Há, portanto, uma gama de definições e compreensões diversas entre autores especializados. Krieger e Finatto (2004) afirmam que a macroestrutura abarca a estrutura global composta por partes introdutórias, as entradas e outras informações. Para Welker (2004, p. 81), macroestrutura "refere-se à forma como o corpo do dicionário é organizado". O autor considera também questões relacionadas com o arranjo das entradas, as características dos verbetes, ou seja, deve-se ou não manter o formato para todos. Outro aspecto relaciona-se à inclusão de tabelas e de ilustrações; por fim, se haverá inserção de informações sintáticas ou outras dentro ou fora do bloco do verbete.

Ao organizar a ordem alfabética das entradas, o lexicógrafo deverá levar em conta alguns critérios que devem ser homogêneos. Por exemplo, ao elaborar as remissivas, adotar critérios minuciosos para que, ao remeter a um termo, não incorra no erro de não haver o correspondente. Em relação a variáveis como palavras flexionadas em gênero, número e grau, prever um critério único de elaboração das entradas que pode ser, por exemplo, adotado o critério de o verbete estar sempre no masculino, singular e grau normal.

A organização macroestrutural deverá prever também questões de tratamento de palavras compostas, fraseologismos, abreviatura e siglas, dentre outros. A decisão ficará a critério de lexicógrafo, contudo deverá seguir um padrão de funcionamento. Seja qual for a decisão, o lexicógrafo deverá levar em consideração o tipo de organização da obra: a organização semasiológica e a onomasiológica. Couto (2012, p. 186) afirma que "a onomasiologia vê a questão da referência, para usar um termo semiótico, partindo da coisa e indo na direção do nome que ela recebe. A semasiologia, no mesmo contexto, faz o percurso inverso, partindo da palavra e indagando a que coisa, ou coisas, ela se refere".

Na organização onomasiológica, o consulente busca o sema ou o verbete a partir de um conceito ou significado já adquirido. Isso quer dizer que, ao se ter em mente uma estrutura

ou uma ideia que se deseja exprimir por uma palavra, o dicionário, com base na organização onomasiológica, daria conta.

Por outro lado, na organização semasiológica, o consulente tem a palavra e a busca entre o verbete no dicionário, para se encontrar o conceito ou o significado. A classificação semasiológica das entradas é a forma tradicional de elaboração dos dicionários impressos. Haensch (1982) ressalta que a ordem alfabética das entradas é o mais importante princípio para ordenação das palavras.

Porém, cabe ressaltar que, quando se trata de uma obra lexicográfica disponibilizada *online*, o procedimento de ordenação em ordem alfabética é desconstruído, já que o usuário, ao fazer uma consulta, na maioria das vezes, digita a palavra no campo de busca e o sistema lhe oferece todas as informações disponibilizadas pelo lexicógrafo. Ao consultar o dicionário semasiológico em ordem alfabética, o consulente tem a oportunidade de observar outros verbetes na mesma página da palavra pesquisada. Isso não seria possível nos dicionários eletrônicos, já que, em sua maioria, disponibilizam para visualização apenas a palavra pesquisada. Porém há outras informações que poderiam ser visualizadas, as quais, por questões dimensionais das obras impressas, não seriam possíveis.

Além da macroestrutura, ao se elaborar uma obra lexicográfica há de se observar também a microestrutura, conforme apontado na seção seguinte.

3.4.2 Microestrutura

A microestrutura de uma obra lexicográfica, em oposição à macroestrutura, refere-se à estrutura interna de um verbete, ou seja, o conjunto de informações que acompanha cada um dos verbetes que compõe a obra lexicográfica. Vilela (1995, p. 233) nos diz que "na microestrutura temos normalmente a biografia da palavra: o seu 'bilhete de identidade', o seu peso e aceitação pela comunidade, a etimologia, a sua idade e posição social, através da caracterização como 'arcaica', 'desusada', 'coloquial', 'familiar', 'gíria', 'calão', etc".

Pode-se dizer então que a microestrutura é responsável pelas informações da palavra. De acordo com García Palacios (2002), a microestrutura é composta por três tipos de informações:

i) as informações gerais em relação à palavra, neste caso o verbete, como separação silábica, indicações fonéticas e etimológicas, além de informações sobre os dados cronológicos, etimologia e variações dialetais.

ii) as informações sobre o funcionamento do verbete, como, por exemplo, aquelas de natureza gramatical como gênero e número, e contextos abonatórios ou frases exemplificando o uso;

iii) as informações sobre o significado, que estão relacionadas com a definição ou com a descrição do verbete.

Além disso, Martins (2017, p. 189), com base em Zavaglia (2012), aponta que a microestrutura pode ser organizada seguindo os critérios:

i) Abonação – uma frase retirada da obra de um autor consagrado, refletindo o bom uso da unidade lexical, ou o uso idiossincrásico do autor;
ii) Exemplo autêntico – frase extraída de um *corpus* composto de diferentes gêneros e tipologias textuais e que representa o uso real da unidade léxica
iii) Exemplo criado, inventado ou forjado – elaborado a partir da intuição do lexicógrafo, refletindo o uso correto, porém não real, do item;
iv) Exemplo adaptado – retirado de um *corpus* e reescrito pelo lexicógrafo.

Ainda sobre a microestrutura dos verbetes, Biderman (2001) afirma que

> a microestrutura tem como eixos básicos a definição da palavra em epígrafe e a ilustração contextual desse mesmo vocábulo, quer através de abonações por contexto realizado na língua escrita ou oral, quer através de exemplos. Quanto à ilustração contextual (e/ou abonação) ela é essencial para explicitar claramente o significado e/ou uso registrado na definição. Os significados e usos referidos são aqueles já registrados e documentados em contextos realizados e não valores semânticos possíveis, eventualmente atribuíveis ao lexema da língua. O verbete deve ser completado com informações sobre registros sociolinguísticos do uso da palavra e remissões a outras unidades do léxico associadas a este lema por meio de redes semântico-lexicais. (BIDERMAN, 2001, p. 18).

Portanto, ao se elaborar a definição de um verbete, deve-se levar em conta de que se trata de um recorte da amplitude do significado, dentro da gama de possibilidades de usos em

uma língua. A definição visa, dessa maneira, a promover e a facilitar interação entre os sujeitos que participam do ato de comunicação.

Cardoso (2017) apresenta o modelo de verbete, proposto por Barbosa (1999), que pode ser constituído da maneira apresentada no Quadro 2.

Quadro 2: Modelo de verbete proposto por Barbosa

Verbete = [+ entrada (lexema) + Enunciado Lexicográfico (+/- **Paradigma Informacional (PI)1** (pronúncia, abreviatura, categoria, gênero, número, etimologia, homônimos, campos léxico-semânticos etc.), + **Paradigma Definicional (PD)** (acepção$_1$, acepção$_2$ ... acepção$_n$) +/- **Parad**. **Pragmático (PP)** (classe contextual$_1$, classe contextual$_2$, ... classe contextual$_n$). +/- PI.2, ... PI.n) +/- Remissivas da cadeia interpretante de língua)].

Fonte: Barbosa (1999 apud CARDOSO, 2017)

Em relação à organização dos verbetes, Borba (2003, p. 322) afirma que, inicialmente, deve ser guiada pela taxionomia, a qual controla o sistema definitório. Afirma também que existem particularidades em relação à natureza e estabelece dois conjuntos: palavras lexicais e palavras gramaticais. As palavras lexicais "organizam-se em matrizes valenciais"; já as palavras gramaticais, "pela natureza das relações que estabelecem".

Para Borba (2003, p. 322), "o verbete mais simples terá três níveis de informação: a classe da palavra a que pertence, a definição ou equivalência sinonímica e a abonação". Para o autor, as acepções ampliam o verbete, porém não lhe acrescentam um nível. Porém, ainda segundo Borba (2003, p. 322), o tipo de verbete mais comum é "aquele que contém quatro ou cinco níveis de informação: a classe, a subclasse, a complementação, a definição e a abonação". Entretanto, há ainda aqueles que podem conter mais dois níveis: "a adjunção de expressões ou frases feitas e observações sobre usos especiais". O autor ainda adverte que

> Cada um desses itens pode ampliar-se, ou seja, uma palavra pode ser usada em mais de uma classe, mais de uma subclasse ou mais de um registro; pode ter mais de um complemento ou mais de uma acepção, a variedade e complexidade de estruturas de verbetes realmente é muito grande. (BORBA, 2003, p. 323)

Entendemos que há propostas de estruturas de verbetes, porém elas devem atender aos propósitos da obra lexicográfica organizada e, como afirma Borba (2003), deve estar apoiada numa teoria gramatical que a sustentará.

Outro aspecto relevante diz respeito à definição. Segundo Barbosa (2004, p. 76-77), a definição é uma "[...] estrutura sintático-semântica, sua forma de conteúdo e expressão, requerida por esse tipo de discurso parafrástico, em que os traços conceptuais são organizados em forma de frase, ou seja, manifestados como metatermos".

Borba (2003) adverte que a seleção dos elementos descritivos que comporão o verbete está relacionada com o objetivo da obra lexicográfica. Isto é, se o objetivo for registrar todo o uso, caberiam informações sobe todos os níveis da estrutura linguística, como informações sobre fonética, morfologia, sintáticas, semânticas e pragmáticas. Após a definição dos elementos, deve-se partir para a elaboração da hierarquia dos níveis. O autor ainda propõe alguns níveis: a primeira informação diz respeito à classe a que pertence a palavra, ou seja, é de ordem taxionômica; a seguir, vem a definição, que varia de acordo com o lexicógrafo e pode ser em forma de conceituação, explicação ou descrição. Na conceituação e explicação são usados termos que ocorrem no *corpus* com frequência maior, ou seja, parte-se daquilo que, provavelmente, o consulente conhece; já na descrição, o lexicógrafo procura reter os traços básicos do item lexical. A descrição, segundo o autor, é comum quando se refere a plantas.

Em razão do exposto, para se criar um verbete, independente da obra lexicográfica a ser elaborada como dicionário, glossário ou vocabulário, como o caso particular desta tese, é necessário definir parâmetros, com os quais se definirão a quantidade e a sequência das informações que serão disponibilizadas. O lexicógrafo tem opções de formulação dessas informações a partir do seu próprio conhecimento da língua ou das línguas para o caso dos dicionários bilíngues e multilíngues, ou buscar informações em outras fontes como nos contextos abonatórios, com especialistas, como também em outras obras lexicográficas.

Expusemos questões relacionadas à constituição de obras lexicográficas. A seguir fazemos uma breve exposição sobre as obras lexicográficas eletrônicas.

3.5 Dicionários eletrônicos

O avanço da tecnologia impactou, sobremaneira, nos trabalhos dos lexicógrafos, no sentido de auxiliar na elaboração, divulgação e armazenamento dos dicionários, dos glossários e dos vocabulários. Nesse sentido, há os dicionários disponíveis em mídias para computadores e os disponíveis *online* para consulta. Welker (2004, p. 225) estabelece que o termo dicionários eletrônicos refere-se a dicionários;[8] "1) usados no processamento computacional da linguagem natural; 2) em CD-ROM; 3) *online* (acessíveis na internet); 4) portáteis." Os dicionários em CD-ROM e os dicionários *online* têm características parecidas, porém distinguem-se pelo fato de que os dicionários em CD-ROM não podem ser atualizados; já os dicionários *online* podem receber novas informações e serem reestruturados a qualquer momento. Outro aspecto é que os dicionários *online* precisam de conexão com a *internet* para acesso, por outro lado, os dicionários em CD-ROM dependem somente de um dispositivo eletrônico para leitura.

Os dicionários eletrônicos, segundo análise de Villalva e Silvestre (2014, p. 165), "distinguem-se dos impressos pelo fato de explorarem hiperligações entre as palavras, por ampliarem a quantidade de dados consultáveis e pela interação com aplicações de correção ortográfica e paradigmas de flexão e conjugação". Além disso, para os autores, os dicionários eletrônicos podem ser explorados como um amplo *corpus* textual, o que o diferencia do dicionário impresso, cujas informações são limitadas. Quebra-se, dessa maneira, o princípio da ordenação alfabética e da procura somente pelo verbete. Há inúmeras possibilidades de busca pelos marcadores metalinguísticos como datação, étimo, domínio lexical ou campos semânticos.

Sobre a distinção entre os dicionários eletrônicos e impressos, Costa (2014) elaborou um quadro pontuando os principais itens. O Quadro 3 apresenta a distinção feita por essa autora.

[8] Empregamos a palavra dicionário, em razão de os autores também o fazerem, porém entendemos a teoria escrita sobre os dicionários dão conta também do que se refere a outras obras lexicográficas como os glossários e vocabulários.

Quadro 3: Características dos dicionários impressos e eletrônicos

Tópico	Impresso	Eletrônico
Forma de constituição	Características físicas palpáveis através da impressão	Formado por ondas transmitidas de um computador a outro, ou de um dispositivo ao computador
Aparição	Permite a visão geral do modelo e de todos os termos lematizados	Possibilidade de ocultar o glossário, aparecendo somente o termo consultado
Consulta	O consulente deve procurar o termo junto aos demais termos que compõem a macro e microestrutura	O consulente digita e tem acesso ao termo que procura
Atualização	Requer nova impressão de todo o glossário	Atualização a qualquer tempo
Custo de produção	Alto	Baixo
Utilização	O consulente deve interromper a leitura para procurar o termo pesquisado	O consulente digita o termo e a informação vem praticamente em ato contínuo
Utilização	O consulente deve interromper a leitura para procurar o termo pesquisado	O consulente digita o termo e a informação vem praticamente em ato contínuo
Tamanho	Não permite compactação	Permite compactação
Recursos	Não requer recursos adicionais para a sua utilização	Requer, como recurso adicional, equipamento eletrônico que possibilite a leitura do *software*
Vantagens	Pode ser lido em qualquer lugar	Somente pode ser acessado em equipamento eletrônico
Peculiaridade	Maior tempo de busca	Redução do tempo de busca

Fonte: Costa (2014, p. 97)

Villalva e Silvestre (2014) também mencionam que os dicionários eletrônicos, que podem ser consultados pela *internet* e estão vinculados a uma base de dados, podem ser

constantemente atualizados e, portanto, podem ser-lhes acrescentados elementos novos que é a perspectiva da constante atualização. Os autores complementam que

> Os alargamentos do *corpus*, a revisão da nominata e as correções repercutem-se mais rapidamente nos resultados disponíveis para o utilizador, o que alterou o estatuto normativo dos dicionários eletrônicos, anteriormente considerados como e pontos de referência para o estudo diacrônico. (VILLALVA; SILVESTRE, 2014, p. 197)

Pela facilidade de consulta e de atualização dos dados, o dicionário eletrônico tende a ser uma preferência dos consulentes, já que se obtêm um maior número de informações sobre um item lexical de uma língua do que o dicionário impresso, além de não ocupar espaço físico com livros enormes em prateleiras.

Leffa (2006, p. 323) também estabelece uma distinção entre os dicionários convencionais, de papel, e os eletrônicos. Segundo o autor, o dicionário eletrônico é armazenado em arquivo digital, por isso extremamente maleável, ou seja, "pode ser ampliado e atualizado, sem grandes custos de produção", como também possibilita a inclusão de animação, som e vídeo. Além disso, Leffa (2006) afirma que outra característica do dicionário eletrônico é o caráter de invisibilidade, por isso pode aparecer somente quando for requisitado pelo consulente, como também poderá ser mostrado apenas o verbete solicitado, portanto todo o resto ficaria oculto no suporte que o sustenta. Já o convencional, o impresso em papel, não há possibilidade de compactação, portanto é um texto de volume vultoso, o que dificultaria o transporte, por exemplo. Além disso, qualquer atualização demandaria nova impressão com altos custos de produção, como também impossibilita a inserção de animações, som e vídeo.

O autor finaliza as diferenças afirmando que a maior delas está no acesso ao verbete desejado pelo leitor. Leffa (2001) afirma que, ao usar o dicionário impresso em papel, a consulta é obstrutiva, ou seja, o leitor interrompe a leitura, move-se para outro texto e inicia outro tipo de leitura, o que, às vezes, demanda um tempo em razão da busca pelas diversas páginas até localizar a palavra que se procura. Então ocorre o processo inverso que é o retorno ao texto original, onde, novamente, localizam-se as partes nas quais ocorreu a interrupção. Então, o ideal seria que o leitor pudesse consultar, em menor tempo, o que se deseja para não interromper o ritmo da leitura "para que o fluxo das ideias necessário para a compreensão do texto não seja interrompido." (LEFFA, 2001, p. 04). Em razão disso, o autor propõe duas características básicas para um dicionário:

- permitir acesso instantâneo ao verbete – o tempo que demandaria ao leitor para consultar uma palavra deveria ser imperceptível para que a interrupção, sendo em tempo mínimo, não prejudicar a construção do sentido.

- estar subordinado ao texto – a consulta ao dicionário não deve afastar o texto da frente do leitor. Isto quer dizer que o dicionário deve ajudar o leitor de maneira discreta, sem estabelecer competição com o texto lido. Isso porque "o significado da palavra não está no dicionário, mas no texto que está sendo lido. O dicionário apenas dá pistas; quem dá o significado da palavra é o texto." (LEFFA, 2001, p. 05).

Em consonância com o autor, um acesso instantâneo, sem o abandono da leitura, pareceria ser mais viável que a consulta em dicionário impresso. No caso do vocabulário do léxico indianista de Alencar, também é essencial a disponibilização eletrônica, para que o leitor mantenha a atenção e concentração no texto, evitando a dispersão decorrente da busca em um dicionário impresso. Outro ponto importante é o fato de que até o momento de finalização desta tese encontramos um total de cinco obras lexicográficas disponibilizadas online: *Dicionário Ilustrado Tupi Guarani*[9], *Dicio, Diconário on line de Português*[10], *Dicionário Indígena*[11], *Minidicionário indígena*[12] e *Minidicionário Tupi*[13]. Especificamente, com o léxico indianista de Alencar não há algum que se tenha conhecimento, até o momento de redação desta tese não foi localizado.

O intuito é o vocabulário eletrônico possa ser consultado a partir de qualquer dispositivo móvel conectado à *internet*. Caso o leitor opte pela leitura dos romances também em formato digitalizado, ele teria um aparato que garantiria a consulta sem ter que abandonar o suporte, o dispositivo móvel, para, depois, retornar a ele para continuar a leitura. Portanto, o dicionário eletrônico pode levar o leitor a compreender mais em menos tempo.

Sobre as possibilidades de busca, Welker (2004, p. 228) afirma que os dicionários eletrônicos apresentam mais facilidade em relação aos impressos: i) caso o usuário não se lembre da palavra inteira, ao digitar uma parte, "recebe como resultado todos os lemas que

[9] Disponível em www.dicionariotupiguarani.com.br. Acesso em: 02 maio 2018.
[10] Disponível em www.dicio.com.br/palavras-indigenas. Acesso em: 02 maio 2018.
[11] Disponível em www.dicionarioindigena.blogspot.com.br. Acesso em: 20 maio 2018.
[12] Disponível em www.cambito.com.br. Acesso em: 02 maio 2018.
[13] Disponível em: https: maniadehistoria.wordpress.com/mini-dicionario-tupi-guarani. Acesso em: 20 maio 2018.

contêm o grupo de letras digitado, o que talvez o ajude a se lembrar da palavra"; ii) há alguns dicionários que são projetados para encontrarem o verbete correto mesmo que o consulente o grafe erroneamente; iii) consulentes também podem, por meio da consulta, obterem listas de palavras pertencentes a determinada classe gramatical, por exemplo: "escolhendo-se gíria e a letra *b*, são mostradas todas as acepções de lexemas com inicial *b* e que contenham essa marca"; iv) outro recurso é a possibilidade de haver *links* dos verbetes para outros dicionários, ou outras informações, principalmente sobre as abonações do verbete e outras informações gramaticais, como conjugação dos verbos.

O dicionário eletrônico pode até ser constituído pelos mesmos elementos do impresso, porém a disposição das informações, tendo em vista o suporte, podem ser apresentadas de forma a diversificar a consulta, por isso o leitor poderá escolher a palavra e as informações que deseja saber sobre ela, por exemplo, significado, etimologia, sinônimos, antônimos, dentre outros. Portanto, o suporte eletrônico não ajudará muito, se as informações lexicais não estiverem disponíveis para acesso.

Continuando o embasamento teórico desta pesquisa, são apresentados, na seção seguinte, os conceitos de semântica e de campo semântico

3.6 Campos semânticos

A identificação de campos semânticos é essencial, já que, em suas obras indianistas, Alencar procurou criar um contexto peculiar dos índios. Assim, analisar o léxico indianista do autor nas obras *Iracema*, O *Guarani* e *Ubirajara* é também estratificar esse léxico em campos semânticos.

Para iniciar, Guirraud (1972, p. 106) define semântica como "tudo o que se refere ao sentido dos signos convencionais por meio dos quais exprimimos ideias com a finalidade de comunicá-las". Para o autor, as palavras são criadas com a finalidade de nomear as coisas por duas razões, ou porque o que se deseja nomear ainda não tem nome ou porque os nomes que elas têm não são eficazes na sua função. Apesar de ser criação individual para nomear as coisas, as palavras são disseminadas de forma coletiva e inconsciente. As palavras são criadas pelo homem, porém adquirem vida própria, ou seja, elas se criam, segundo Guirraud (1972, p.

45) “uma vez criada a palavra, por transferência de sentido ou por qualquer outro modo, seu sentido pode evoluir espontaneamente; de fato, na quase totalidade dos casos ele evolui”.

No intento de analisar as palavras, Cançado (2005) afirma que

> a semântica pode ser pensada como a explicação de aspectos da interpretação que dependem exclusivamente do sistema da língua e não de como as pessoas o colocam em uso: em outros termos, podemos dizer que a semântica lida com a interpretação das expressões linguísticas, como o que permanece constante quando uma certa expressão é proferida. (CANÇADO, 2005, p. 17).

A autora concebe a interpretação de uma palavra considerando o sistema, ou seja, a palavra em uso. Ampliando, Cançado (2005, p. 18) afirma que a semântica “não pode ser estudada somente como a interpretação de um sistema abstrato, mas também tem que ser estudada como um sistema que interage com outros sistemas, no processo de comunicação e expressão dos pensamentos humanos.”

As palavras, então, não podem ser estudadas como parte de um sistema abstrato, mas num sistema em movimento, pois os sentidos estão ligados intimamente com o contexto de produção. Nesse sentido, Biderman (2001) afirma que o “universo semântico se estrutura em torno de dois polos opostos: o indivíduo e a sociedade” e, partir dessa tensão em movimento, se origina o léxico.

Considerando o universo semântico, vale ressaltar que as palavras podem ser organizadas em campos semânticos. Biderman (2001, p. 193) afirma que os “campos semânticos podem evidenciar oposições simples, e/ou oposições complexas de significação. Entre as oposições simples podemos incluir aquelas em que os termos integrantes do campo semântico distinguem-se apenas por um dou dois traços semânticos”. Segundo a autora, um grupo como palácio, palacete, mansão, casa, casinha, choupana e casebre carregam os componentes sêmicos de “casa”, porém há os traços de significação que os distinguem que são tamanho e riqueza/pobreza.

Para Guirraud (1972, p. 104) o campo semântico é “o conjunto de relações do qual cada termo tira sua motivação, mas de relações não necessárias e não sistemáticas. Esse caráter contingente das relações léxicas parece proibir qualquer esperança de se considerar o léxico como um sistema inteiramente estruturado”. Essa noção de o léxico ser um sistema que

não se estrutura vai ao encontro do que autores afirmam em relação ao léxico de uma língua ser um sistema aberto e em constante expansão.

Biderman (2001) ainda chama a atenção para um aspecto relevante na análise dos campos semânticos em que se estrutura o léxico: às palavras empregadas nos diversos contextos, podem ser adicionadas conotações diferentes em decorrência das combinações sintáticas dos sintagmas. Nesses contextos, novos matizes de significação são adicionados às palavras decorrentes da combinatória de elementos diversos do léxico, sendo assim, ressalta a autora (p. 198), "o lexicólogo não poderá delimitar o campo de significação de uma palavra, pois o léxico engloba todo o universo da significação, o que inclui toda a nomenclatura e interpretação da realidade".

Pode-se atestar o exposto em Cançado (2005, p. 17) ao afirmar que "nem sempre o sistema semântico é o único responsável pelo significado; ao contrário, em várias situações, o sistema semântico tem o seu significado alterado por outros sistemas cognitivos para uma compreensão final do significado". Guirraud (1972, p. 64) ressalta que "o sentido muda porque se dá deliberadamente um nome a um conceito para fins cognitivos ou expressivos; porque as coisas são nomeadas. O sentido muda porque uma das associações é secundária (sentido contextual, valor expressivo, valor social); ele desliza progressivamente sobre o sentido de base que o substitui; o sentido evolui".

Por esse mesmo viés, Cançado (2005, p. 19) afirma que "uma teoria semântica deve, em relação a qualquer língua, ser capaz de atribuir a cada palavra e a cada sentença o significado (ou significados) que lhe (s) é (são) associado (s) nessa língua. No caso das palavras, isso significa essencialmente escrever um dicionário".

Sobre a composição dos dicionários, Katz e Fodor (1977, p. 97) afirmam que do ponto de vista de uma teoria semântica, um verbete de dicionário é constituído de duas partes: "uma parte gramatical, que fornece a classificação relativa às partes do discurso para o item lexical, e uma parte semântica, que representa cada um dos distintos sentidos que o item lexical possui em suas ocorrências como uma dada parte do discurso". A estrutura lexicográfica torna bastante complexa no que se refere, principalmente, à parte semântica, pois, além das imagens e analogias externas entre as coisas denominadas, as aproximações e contaminações de sentido desempenham um papel importante na estruturação do termo.

Essas aproximações de sentido, muitas vezes, são percebidas pelos falantes de uma língua em razão de uma certa intuição e da dedução que, por vezes, o próprio falante realiza. Cançado (2008) afirma que

> os falantes nativos de uma língua têm algumas intuições sobre as propriedades de sentenças e de palavras e a maneira como essas sentenças e palavras se relacionam. Por exemplo, se um falante sabe o significado de uma determinada sentença intuitivamente sabe deduzir várias outras sentenças verdadeiras a partir da primeira. Essas intuições parecem refletir o conhecimento semântico que o falante tem. Esse comportamento linguístico é mais uma prova de que seu conhecimento sobre o significado não é uma lista de sentenças, mas um sistema complexo, ou seja, o falante de uma língua, mesmo sem ter consciência, tem um conhecimento sistemático da língua que lhe permite fazer operações de natureza bastante complexa. (CANÇADO, 2008, p. 20).

O comportamento linguístico não é regido por regras capazes de identificar todas as possibilidades de utilização e uma palavra e/ou de todas as sentenças de uma língua. Assim sendo, ratificando Cançado (2008), o falante cria e recria palavras e significados de acordo com as necessidades do momento.

Pottier (1972) propõe que haja uma semântica da frase, ou seja, uma semântica na qual o contexto é que determinará o valor do vocábulo, o que excluiria a semântica da palavra. Assim sendo, entende-se que é necessária uma análise em que se inter-relacione o vocábulo com o contexto, imprescindível para se abstrair os traços semânticos do vocábulo. Para Pottier (1972, p. 42), por mais que um dicionário procure enumerar todos os domínios, "a noção de significação é sempre relativa e supõe uma situação de discurso que atualiza o domínio".

Na seção seguinte, cumpre-nos prestar alguns esclarecimentos em relação aos antropônimos e aos topônimos.

3.7 Antropônimos e topônimos: algumas considerações

Não é comum que os dicionários insiram os antropônimos ou topônimos entre os seus verbetes, ressalvando os especializados. No caso dos romances indianistas de Alencar, essas categorias cumprem um papel de "fio condutor" que nos leva ao conhecimento do cenário onde as narrativas ocorrem, quando se trata dos topônimos. Já em relação aos antropônimos,

Carvalhinhos (2007, p. 02) nos informa que "a grande diferença é que no começo dos tempos [...] o nome era conotativo, isto é, sua carga significativa era perfeitamente decodificável", ratifica, portanto que, no caso dos romances indianistas de Alencar, os nomes exercem um papel central na constituição da narrativa, porque não apenas nomeia, mas imprime características aos personagens.

As ciências que estudam os antropônimos e os topônimos são, respectivamente, a Antroponímia e a Toponímia, ambas, por sua vez, compõem a ciência que estuda a formação dos nomes próprios, a Onomástica, um ramo da Lexicologia. Os nomes próprios de pessoas, incluindo os sobrenomes (ou nomes parentais) e os apelidos são objeto de estudo da Antroponímia. A Toponímia, por sua vez, investiga as motivações dos nomes de lugares.

O estudo da Onomástica é relevante de tal modo que Dick (1999) afirmou que as duas ciências, a Antroponímia e a Sinonímia,

> ultrapassam, em muito a conceituação teórica que lhes é atribuída, tornando-se Ciências Humana, fontes de conhecimento tão excelentes quanto as melhores evidências documentais. São por assim dizer, verdadeiros registros do cotidiano, manifestado nas atitudes e posturas sociais que, em certas circunstâncias, a não ser deles, escaparia às gerações futuras. (DICK, 1999, p. 178)

Assim sendo, o estudo do léxico indianista, visto que é perpassado por antropônimos e topônimos, é imprescindível para resgatar a própria história da Língua Portuguesa, de caráter diacrônico, como estudar o contexto em que Alencar estava inserido e as suas obras, de maneira particular. A nomeação tem relação direta com aspectos culturais, histórico e sociais de um povo, portanto podem acumular e conservar informações sobre épocas quaisquer. No caso do inventário lexical de Alencar, podemos constatar que inúmeros elementos da vida cultural dos indígenas estão refletidos tanto no se refere aos antropônimos e topônimos, como no acervo lexical de maneira geral.

Adotamos a definição de antropônimo segundo Amaral (2008, p. 70) de que "item lexical que, em um contexto determinado, nomeia um indivíduo ou é utilizado para fazer referência a um indivíduo do mundo real ou fictício". Assim sendo, comporão o campo semântico dos antropônimos os nomes próprios e os nomes que se referem aos indivíduos, a exemplo de 'criança' na língua portuguesa.

Outro campo semântico depreendido pela análise dos vocábulos indianistas são os topônimos que, segundo Isquerdo

> um topônimo além de determinar a identidade de um lugar, a análise e sua estrutura pode fornecer elementos para esclarecer muitos aspectos referentes à história política, econômica e sociocultural de uma região. Desta forma, o papel do signo toponímico ultrapassa o nível apenas da identificação, servindo, pois, de referência para o entendimento de aspectos da realidade em que está inserido. (ISQUERDO, 1997, p. 9).

No caso dos indígenas, a observação da Isquerdo (1997) calha, pois os índios nomeavam os elementos referentes ao lugar baseando-se nas características desse lugar, o que facilitava a compreensão, identificação, localização e a descrição do elemento geográfico descrito.

Não pretendemos esgotar as teorias sobre Antroponímia e Toponímia, mas apenas apresentar alguns apontamentos que justificam a manutenção desses campos semânticos em nosso vocabulário com as respectivas descrições.

Na próxima seção, trazemos as questões teóricas relacionadas à Etimologia que compõem o arcabouço teórico que sustentam esta tese.

3.8 Etimologia

As palavras fazem parte do léxico de uma língua, porém não de forma arbitrária. O léxico vai se formando a partir de uma evolução caracterizada por empréstimos, por criação de novas palavras ou por ressignificação das já existentes. O conhecimento da origem das palavras é importante, porque também é uma forma de conhecer a própria história de uma sociedade. Entretanto, descobrir a origem das palavras de uma língua não é uma tarefa fácil, às vezes, afirmar a verdadeira origem de determinada palavra é impossível, dada a dinamicidade de uma língua viva.

Não obstante a essa dificuldade, a etimologia se ocupa de saber não só a respeito da origem da palavra, mas também do seu significado atual, sua história e seus processos de formação. Etimologia é, então, o estudo da origem das palavras, juntamente com as sucessivas mudanças pelas quais essa palavra passou (CASARES, 1992). As mudanças podem ser de diversos níveis como no sentido, no som ou na grafia, como, por exemplo, palavras do latim

como **periculum**, **nuptiae** e **bonus** que deram origem, respectivamente, a **perigo**, **núpcias** e **bom,** em Língua Portuguesa atual. São inúmeros os exemplos de casos de mudanças pelas quais os vocábulos utilizados, atualmente, passaram no decorrer dos anos.

Sobre a importância da etimologia, Casares (1992) afirma que, no início dos estudos etimológicos, procedeu-se à decomposição das palavras em suas raízes e em seus elementos secundários. Esse trabalho possibilitou estabelecer o parentesco de muitas línguas modernas entre si e sua descendência comum em relação a outras línguas antigas. Com isso, foi possível chegar à reconstrução hipotética de línguas e culturas desaparecidas, com base nos vestígios que as línguas dos tempos históricos preservaram nas atuais. Ainda segundo o autor, esta aproximação se deveu muito aos estudos da fonética histórica.

Casares (1992) adverte que, apesar desses procedimentos terem sido bastante frutíferos, eles não poderiam cobrir todo o campo de pesquisa etimológica sozinhos. Os dados oriundos da análise fonética não ofereciam provas suficientes, em muitos casos, para estabelecer definitivamente uma etimologia. Para o autor, além da possibilidade fisiológica do processo formal proposto e da probabilidade psicológica de certas mutações semânticas, era necessário atentar para considerações geográficas, históricas, culturais e até comerciais. Isto porque certos fenômenos linguísticos foram identificados como característicos de uma dada área ou de algumas regiões.

Portanto, não bastava estudar somente os processos formais e semânticos das palavras, ou seja, seria necessário também estudar a sua origem e as suas sucessivas transformações, o que proporcionou o que Casares (1992, p. 31) chama de "etimologia integral ao ar livre"[14]. Esse estudo implica deixar os ambientes de estudo em material impresso, para ir onde a palavra é usada, ou seja, na comunidade de fala. Esse movimento proporcionou a percepção de que as palavras relacionadas até então por sua forma, dentro de determinada área linguística estudada, poderiam se relacionar com coisas diversas como animal doméstico, inseto, instrumento de trabalho ou uma tarefa agrícola, por exemplo. Algumas palavras possuíam um étimo comum, outros eram distintos dos já elencados. A fim de esclarecimentos sobre o vocábulo étimo, Houaiss (2009, s/p) afirma que o étimo é

> 1) termo determinado e abonado (com exceção das formas hipotéticas), que serve de base para a formação de uma palavra; pode ser uma forma antiga (do mesmo idioma ou de outro) de que se origina a forma recente; pose ser o radical com um afixo, pode ser uma palavra moderna a partir

[14] Tradução nossa "etimologia integral y al aire libre"

> da qual se formam outras, pode ser uma forma hipotética (da mesma língua ou de outra) estabelecida para explicar formas recentes; 2) morfema ou palavra que serve de base para a formação de palavras por derivação ou composição; 3) origem de uma palavra; etimologia.

De acordo com o autor, étimo pode ser sinônimo de etimologia, porém há de se considerar o contexto de emprego dos dois termos como tal. Dubois (2007) também traz uma definição para étimo

> é qualquer forma dada ou estabelecida de que se pode derivar uma palavra; o étimo pode ser radical, base a partir da qual se criou, com um afixo, uma palavra recente (...). O étimo também pode ser a forma antiga de que se origina uma forma recente (...). Enfim o étimo pode ser a forma hipotética ou a raiz estabelecida para explicar uma ou várias formas modernas da mesma língua ou de línguas diferentes. (DUBOIS, 2007, p. 251)

O étimo pode ser a forma utilizada para se explicar as palavras de uma língua em uso e a etimologia busca explicar a palavra nos diversos aspectos, ou seja, é uma análise macro no sentido de que se preocupa com questões de formação e de semântica, por exemplo. Houaiss (2009, s/p) propõe a definição para etimologia em três aspectos:

> 1) O estudo da origem e evolução das palavras; 2) a disciplina que trata da descrição de uma palavra em diferentes estados de língua anteriores, até remontar ao étimo; 3) origem de um termo, quer na forma mais antiga conhecida, quer em alguma etapa de sua evolução, étimo.

Já Saussure (1975) diz que etimologia

> É a explicação das palavras pela pesquisa de suas relações com outras palavras. (...) ela faz a história de famílias de palavras, assim como a faz dos elementos formativos, prefixos, sufixos etc. [...] ela descreve fatos, toma emprestados seus elementos de formação tanto à fonética como à morfologia, à semântica etc. Para alcançar seus fins, serve-se de todos os meios que a Linguística lhe põe à disposição, mas não detém sua atenção na natureza das operações que está obrigada a levar a cabo. (SAUSSURE, 1975, p. 250)

O que propõe Houaiss no aspecto três vai ao encontro do estudo dos étimos de Alencar, pois trataremos dos vocábulos indianistas naquilo que consideramos mais antigo, que são os romances indianistas. Isto quer dizer que, com base nos dicionários anteriores à

publicação dos romances e no *Corpus do Português* (DAVIES, 2016), as palavras são consideradas étimos ou não. [15]

Na mesma linha de pensamento de Casares (1992), Viaro (2014) também chama a atenção para o fato de que os estudos etimológicos não podem ser baseados em imaginação e conhecimento de uma língua materna ou de algumas línguas. O autor ainda adverte que

> A pesquisa etimológica, como uma edição crítica, deve passar por muitas etapas rigorosas e, mesmo assim, as soluções de étimo são múltiplas e sujeitas à revisão. A situação, perante uma profusão de étimos (quando bons e dignos de avaliação) é apresenta-los sem uma solução definitiva, da mesma forma que muitas ciências o fazem seriamente com hipóteses não excludentes. (VIARO, 2014, p. 97).

O fato de o autor considerar as hipóteses não excludentes se deve à evidência de que outros pesquisadores podem confirmar ou refutar as hipóteses, por meio de novos dados e argumentos bem fundamentados. O autor ainda ressalta que "não se pode provar uma etimologia apenas por meio da semelhança formal entre o étimo proposto e as palavras investigadas" (VIARO, 2014, p. 98), pois essa semelhança, ainda de acordo com o autor, pode estar relacionada i) à coincidência entre as formas de grafar; ii) pode ser um tipo de empréstimo por razão do contato direto entre línguas; iii) ou serem de origem comum.

Da mesma forma, Miranda (2004) afirma que o estudo etimológico ajuda a saber de onde vem uma palavra, assim como tem a função de oferecer informações sobre as possíveis mudanças no significado de um vocábulo e sobre as alterações morfológicas do étimo, além de indicar a "idade" da palavra.

Casares (1992) ainda ressalta que o etimólogo, apesar de modesto seu campo de ação, ou seja, mesmo que ele pretenda explorar apenas uma língua, ele precisa conhecer o vocabulário das línguas irmãs, conhecer a gramática histórica de cada uma e as peculiaridades fonéticas dos estágios evolutivos. O etimólogo também deve conhecer as relações das línguas com o tronco comum entre elas e ser capaz de localizar no tempo e no espaço os contatos e trocas de atividades e culturas de povos que usam estas línguas.

Portanto, parece consenso entre os autores sobre a dificuldade de se estabelecer um étimo com dada certeza, por isso Viaro (2014. p. 101) afirma que "étimo é um dado de difícil

[15] Os detalhes sobre os dicionários de consulta e o *Corpus do Português* (DAVIES, 2016) estão pormenorizados no capítulo que se destina à Metodologia da pesquisa.

rastreamento, como todo fenômeno linguístico". Em razão dessa dificuldade, o autor menciona que os *corpora* são de grande importância, principalmente aqueles organizados a partir de textos dos quais se podem obter informações de caráter diacrônico. Como fonte de pesquisa, o autor cita alguns *corpora* como o *Corpus do Português*; o *Corpus Histórico do Português Tycho Brahe*; o *Corpus Informatizado do Português Medieval*; o *Corpus de Referência do Português Contemporâneo*; a *Linguateca* e o *Corpus Lexicográfico do Português*.

Além dos *corpora* disponíveis *online*, Viaro (2014) aponta, para o estudo etimológico, os seguintes dicionários: Constâncio (1836); Coelho (1890); Cortesão (1900-1901); Bastos (1928); Nascentes (1932); Machado (1952-1977); Bueno (1963) Guérios (1979); Cunha (1982, 1989, 2006); Fonseca (2001) e Houaiss & Villar (2001).

Estas fontes, dentre outras, são um bom aparato para os estudos etimológicos, porém o que se encontra escrito nelas não garante a etimologia definitiva de um vocábulo, tendo em vista que elas, primeiramente, dispõem de textos disponibilizados após a invenção da imprensa e, em segundo, pelo fato de que não se tem a certeza de que todos os textos estarão arrolados nos *corpora*. Por fim, textos orais de épocas antigas não foram compilados, o que poderá ter acarretado na perda de palavra com as alterações de uma determinada língua. Como o próprio Viaro (2014, p. 102) nos alerta que "as respostas não estão prontas: os autores discordam entre si, propõem várias soluções, elegem esta ou aquela solução e, não raro, erram. (...) porém os *corpora* podem ser os únicos índices para se responder questões sobre a influência de uma palavra sobre outra (...)". Desse modo, está justificada a pertinência quanto á utilização do *Corpus do Português* (DAVIES, 2016), em sua versão histórica e diacrônica, na presente pesquisa.

Essa afirmação vai ao encontro desta pesquisa, uma vez que, por se tratar de vocábulos indígenas, há discordância entre os autores que estudam esta língua, pelo fato de as línguas indígenas serem ágrafas e, portanto, são analisadas tendo como base a aproximação fonética com a Língua Portuguesa. Portanto, os estudiosos, grafaram os fonemas e as palavras de acordo com o que percebiam por meio da audição das palavras ditas pelas bocas dos indígenas.

Outro fato é a utilização do *Corpus do Português* (DAVIES, 2016), em sua versão histórica e diacrônica, que nos proporciona uma das possibilidades de contato com os textos mais antigos escritos em Língua Portuguesa.

Para se estabelecer o étimo de uma palavra, portanto, é necessário ter os *corpora* datados. Porém, conhecer a data da criação de uma palavra é praticamente impossível, mas a datação da ocorrência mais antiga é um limite importante para se saber que, em termos daquela sincronia, a palavra já era usada (VIARO, 2014). É o caso dos étimos alencarianos que são empregados pela primeira vez em textos escritos na Língua Portuguesa pelo autor em seus romances indianistas. Dados consultados no *Corpus do Português* (DAVIES, 2016) e em dicionários anteriores à publicação dos romances, comprovam que alguns vocábulos foram empregados, ou pelo menos publicados pela primeira vez em língua portuguesa, por Alencar. Esses *corpora* de consulta são tão importantes para a etimologia que Viaro (2014) os coloca, em termos de importância, no mesmo patamar dos dados coletados pela Arqueologia ou pela Paleontologia.

Um aspecto semelhante entre os autores, é que eles se referem à forma hipotética sobre determinada palavra, em razão da impossibilidade de se identificar a data precisa da criação de uma determinada palavra. Assim, Dubois (2007, p. 251-252) afirma que "etimologia é a pesquisa das relações que uma palavra mantém com outra unidade mais antiga, de que se origina".

Em razão do exposto, estudar os étimos alencarianos é um campo frutuoso dada a diversidade de vocábulos indianistas que o autor emprega, como também considerando a sua vasta criatividade ao inventar nomes. A seguir, traçamos um panorama sobre etimologia ficcional contextual.

3.8.1 Etimologia Ficcional Contextual

Antes de iniciarmos nossas considerações sobre etimologia ficcional contextual, atemo-nos à etimologia popular. Esse fenômeno é comum em todas as línguas e ocorre quando os falantes ouvem uma palavra ou expressão que não lhes são familiares e as associam com outra já conhecida e familiar. O resultado dessa associação é uma nova palavra que, às vezes, se expande e adquire o universo dos falantes, em alguns casos, expandindo também para a esfera da língua formal. Reportamo-nos a Saussure que afirma que, às vezes, os falantes deturpam as palavras cujos sentidos são desconhecidos, gerando deformações de algumas palavras. Para Saussure (1975, p. 202) "essas inovações, por mais extravagantes que sejam, não se fazem completamente ao acaso, são tentativas de explicar aproximativamente

uma palavra embaraçante relacionando-a com algo conhecido". A esse fenômeno dá-se o nome de etimologia popular.

Segundo Pereira (2014, p. 134), Saussure apresenta dois casos de etimologia popular. O primeiro, é aquele em que a palavra é ressignificada sem que se altere a forma; o segundo, diz respeito à "deformação da palavra para acomodá-la aos elementos que se acreditam conhecer nela". Continua a autora afirmando que a "etimologia popular se reduz a uma interpretação da forma antiga que, mesmo não muito clara, a sua interpretação é o ponto de partida da deformação sofrida. Age em condições particulares e não atinge senão as palavras raras, técnicas ou estrangeiras que os indivíduos assimilam de modo imperfeito". (PEREIRA, 2014, p. 134).

Se na etimologia popular o falante faz associações para identificar a origem da palavra, no caso de Alencar é possível falar de etimologia ficcional contextual, já que o autor buscou explicar os vocábulos empregados nos romances, para os quais não encontrou explicação nos estudos realizados nos livros e dicionários. As razões para as criações dos étimos ficcionais contextuais se deve ao fato de que o autor se baseia em associações de sentido, muitas vezes, a partir da decomposição de palavras oriundas de diversas fontes de pesquisa como Aires de Casal (1754?-1821?), Varnhagen (1816), Gonçalves Dias (1858), dentre outros.

A etimologia ficcional contextual pode atuar tanto na forma quanto no significado das palavras, ou seja, o autor poderá ressignificar uma palavra por aproximação de sentido ou por aproximação fonética, quanto poderá compor uma nova palavra a partir de elementos já conhecidos.

Segundo Pereira (2014, p. 141) "a etimologia popular se insere no processo da analogia, que explica as mudanças de forma dos vocábulos pela interferência dos valores mórficos e semânticos na evolução fonética". Semelhantemente, ocorre com a etimologia ficcional contextual, em que o autor, também por analogia, explica as mudanças ou cria vocábulos por meio do contexto de uso no interior do texto, ou seja, naquele ambiente da narrativa do romance.

Na etimologia ficcional contextual, o autor, por associações conscientes revela a essência da palavra no contexto da obra produzida. No caso de Alencar, o autor procura explicar os vocábulos indígenas criados por meio das notas explicativas ao final dos romances

e das frases explicativas dos vocábulos, ao decorrer da narrativa, buscando, assim, tornar os étimos compreensíveis para os leitores.

Consideramos etimologia ficcional contextual, primeiro, porque alguns vocábulos podem ser considerados étimos do autor; segundo, pois pode ser um vocábulo utilizado para designar narrativas imaginárias ou referir-se a obras criadas a partir de "elementos imaginários calcados no real e/ou de elementos da realidade inseridos em contextos imaginários" (HOUAISS, 2009), como também pelo fato de tratar da criação e emprego de um vocábulo no interior de uma obra de ficção. Por fim, trata-se de contextual, pois conforme Borba (2003, p. 139) "as palavras só se realizam dentro de um contexto. Isso quer dizer que os significados interagem quando combinados por meio de uma relação gramatical. É comum dizer-se ainda o signo, isolado, é opaco, só se tornando transparente quando inserido num contexto".

Borba (2003, p. 141) ainda expande as explicações sobre contexto afirmando que

> contextualidade pode ser entendido como a possibilidade que tem um item de entrar em contexto e contextualização como a própria mecânica de entrada do item em contexto. Enquanto a primeira é estática e potencial, a segunda é dinâmica e atual.(...) se cada item léxico comporta um conjunto (virtual) de traços, é nos contextos que esses traços e combinam, se compatíveis. Não há item sem contexto, mas há muitos deles de contexto único.

Já em relação à contextualização, Borba (2003, p. 143) explica que ela se liga "aos mecanismos que diversificam os itens, vamos dizer que ela se realiza por aquilo que alguns especialistas chamam colocação, ou seja, o conjunto de posições ocupadas por um item a partir de suas exigências básicas".

Consideramos contexto, no caso desta pesquisa, todos os aspectos que envolvem o uso do vocábulo, como exemplo, os nomes escolhidos pelo autor para nomear seus personagens que carregam a carga semântica semelhante às características dos personagens, como **Andira**, guerreiro valente, cujo nome é também de um morcego hematófago. Morcego hematófago se alimenta de sangue e **Andira**, o guerreiro, metaforicamente, se alimenta do sangue do inimigo jorrado durante a guerra

Etimologia Ficcional Contextual é, então, a análise ou a busca da origem dos vocábulos a partir da interpretação no contexto de emprego nas obras indianistas de Alencar. Portanto não temos pretensão de tratar os vocábulos em sua formação na história da Língua

Portuguesa, que cabe ao Etimólogo, mas tratar da etimologia dos vocábulos na evolução dos romances de Alencar, ou seja, baseamo-nos na formação das palavras na história das obras. Por fim, ressaltamos que não é propósito desta tese pesquisar os étimos dos étimos tupis. Por exemplo, não foi objeto de pesquisa buscar qual o étimo de *ceme*.

Na próxima seção, serão apresentados os conceitos de *corpus* e de Linguística de C*orpus* que são delineadores metodológicos e de abordagem nesta pesquisa.

3.9 Linguística de *Corpus*, *corpus*, chavicidade e ferramentas do *WST*: definições e características

O objetivo desta seção é tratar de assuntos relacionados à Linguística de *Corpus* e ao *corpus,* no que se refere às definições e às características, bem como expor algumas questões relacionadas a algumas ferramentas do *WST*: *WordList*, *KeyWords* e *Concord*. A importância dessas colocações reside no fato de que tais temas complementam a base teórica, configurando, assim, um ponto de partida desta pesquisa.

3.9.1 Linguística de Corpus *e* Corpus

Ao termo *corpus* são atribuídas diversas definições, dentre elas, na perspectiva da LC, poderia significar qualquer coleção de textos organizados de acordo com determinado padrão, digitalizados ou não. Berber Sardinha (2004) afirma que "o *corpus* é um artefato produzido para a pesquisa. Assim, embora os textos devam ser naturais (autênticos e independentes do *corpus*), o *corpus* em si é artificial, um objeto criado com fins específicos de pesquisa" (BERBER SARDINHA, 2004, p. 17). Textos naturais são textos produzidos por humanos e que, portanto, existem na linguagem sem o objetivo específico de comporem um determinado *corpus*.

Com o avanço dos estudos, o termo *corpus* também passou por alterações de definição e de abrangência e passou a incorporar também textos orais da língua, além dos escritos, o que possibilitou os estudos da fala, por isso, são comuns os *corpora* de fala de determinados grupos. Berber Sardinha (2004), entretanto, destaca que nem todo conjunto de textos pode ser

considerado um *corpus* e cita, como exemplo, os textos eletrônicos, ou seja, criados a partir de programas de geração de textos.

Apesar de formado por textos naturais, o *corpus* não pode ser coletado de forma aleatória, isso implica dizer que ele deve seguir critérios linguísticos de seleção para seu planejamento e concretização. Segundo Berber Sardinha (2004), o *corpus* deve ser uma coletânea criteriosa, a fim de refletir o mais fiel possível a variante escolhida, além de atender aos objetivos da pesquisa. Por isso, na compilação do *corpus* de estudo, o material textual deve ser apenas o necessário para representar a amostra desejada.

Parodi (2008) apresenta três aspectos relevantes em relação ao *corpus*: i) deve ser composto por textos produzidos em situações reais; ii) a captação das instâncias da língua em uso deve estar guiada por parâmetros explícitos que permitem ter clareza na sua constituição, de modo que se apoiem em análise e metodologias, de tal forma que seja possível replicar em estudos posteriores; iii) um *corpus* deve estar disponível em formato eletrônico, com o fim de ser analisado por programas de computador

Complementando, Berber Sardinha (2000, p. 338-339) sintetiza que há quatro pré-requisitos para a formação de um *corpus* computadorizado: i) o *corpus* deve ser composto de textos autênticos, em linguagem natural; ii) serem escritos por falantes nativos; ii) escolha criteriosa do conteúdo, cujos critérios devem ser a naturalidade e a autenticidade; iv) deve ser representativo de uma variedade linguística ou mesmo de um idioma.

Embora se façam referências a critérios linguísticos de seleção de *corpus* e à sua extensão, não há critérios específicos em relação à extensão do *corpus*, porém ao tratar do tamanho dos *corpora*, Berber Sardinha (1999) traz uma definição de Sánchez que amplia as ideias já expostas sobre *corpus*:

> um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (SÁNCHEZ, 1995 apud BERBER SARDINHA, 1999, p. 12).

Assim, duas questões são colocadas em pauta: a extensão e o processamento do *corpus*. Em relação à extensão, Parodi (2008) afirma que, embora um *corpus* possa ser

formado por, pelo menos, dois ou mais textos, dependendo dos objetivos de pesquisa, quanto maior a proporção de extensão do *corpus* maior também será a probabilidade de ocorrerem os itens a serem pesquisados. Para o mesmo autor, *corpus* corresponde a um conjunto amplo de textos digitais de natureza específica e que conta com uma organização predeterminada em torno de categorias identificadas para descrição e análise de uma variedade da língua.

Também sobre a extensão dos *corpora*, Berber Sardinha (2004) enfatiza que não há critérios definidos em relação ao mínimo da extensão para que um *corpus* seja considerado representativo, porém apresenta três abordagens a se considerar na composição do *corpus*. A primeira, chamada de impressionista, que se baseia em observações e constatações de especialistas da área sobre a criação e exploração dos *corpora*. A segunda, considerada, pelo autor, de histórica, e leva em conta a monitoração dos *corpora* efetivamente utilizados por uma comunidade, em um determinado período. A última abordagem é a estatística, que se fundamenta em dados estatísticos, para definir a quantidade de palavras que seriam necessárias para se constituir uma amostra capaz de representar determinados aspectos ou características de uma língua.

Considerando-se as questões relacionadas à extensão do *corpus*, Berber Sardinha (2000) elaborou uma tabela que, na sua perspectiva, sintetiza a classificação dos *copora* em relação ao tamanho em palavras:

Tabela 1: Classificação do *corpus* em relação à quantidade de palavras

Tamanho em palavras	Classificação
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Fonte: Berber Sardinha (2000, p. 346)[16]

Parodi (2008) traz as recomendações de Eagles (1996) sobre a constituição do *corpus*, para que ele possa ser chamado como tal: i) o *corpus* deve ser o mais extenso possível, de

[16] Não foram localizados dados atualizados sobre o tamanho de *corpora* com base na quantidade de palavras.

acordo com as tecnologias disponíveis; ii) deve incluir exemplos de vários materiais, para ser o mais representativo possível; iii) deve haver uma classificação intermediária dos gêneros, em relação ao *corpus* total e a amostras individuais; iv) as amostras devem ter tamanhos semelhantes; v) o *corpus*, como um todo, deve ter procedência clara.

Em relação ao processamento do *corpus*, Berber Sardinha (2004) afirma que o computador é a "mola propulsora" responsável pela revolução e pelos avanços nos estudos linguísticos. E afirma, também, que "para entender essa revolução, é preciso acompanhar a Linguística de *Corpus*, uma área que trata do uso de *corpora* computadorizado". (BERBER SARDINHA, 2004, p. 2).

Dessa forma, o computador possibilitou realizar pesquisa e processar informações complexas em bancos de dados de tamanha extensão que a mente humana, sem o apoio de uma máquina, seria incapaz de realizar. Um exemplo, seria a elaboração de dicionários. Assim, a LC se apresenta como um novo caminho para os linguistas, professores, tradutores, lexicógrafos e outro profissionais (BERBER SARDINHA, 2004). Na década de 1960, o trabalho era realizado manualmente, ou seja, as palavras eram transferidas manualmente para cartões, a fim de que se pudesse realizar a leitura por programas de computador, o que apresentava dificuldades advindas dos recursos humanos e de tempo.

A LC, por meio de ferramentas computacionais, auxilia na pesquisa de uma ou mais línguas por meio de observação e descrição de grandes quantidades de textos digitalizados. A LC, então, permite entender os textos como um sistema probabilístico de ocorrências, por meio de padrões lexicogramaticais que avaliam a palavra e as situações reais de ocorrências em que ocorrem de fato. LC, na definição de Berber Sardinha,

> ocupa-se da coleta e exploração de *corpora*, ou conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador. (BERBER SARDINHA, 2004, p.3).

Por meio do computador, é possível, então, ter acesso ao contexto real de uso de uma língua, estudar o que é escrito ou falado. Porém, Parodi (2008) adverte que é um *corpus* de uma variedade linguística não representa a língua na qual está inserido, se considerarmos a diversidade e variedade de cada língua em particular. Assim, o *corpus* é somente uma coleção finita de um universo infinito. Pode-se afirmar, então, que um *corpus* oferece informações

sobre uma língua em particular, porém, pode-se dizer que é impossível coletar um *corpus* que represente toda uma língua.

A LC está voltada para a compilação de textos para que pesquisadores interessados nos fatos da língua possam elucidar os aspectos propostos no estudo. Por meio da LC é possível fornecer descrições e explicações acerca de elementos linguísticos de quaisquer ordens: lexical, gramatical como fonológico, morfológico, sintático e semântico. Assim, o *corpus* não se resume a coleta de *corpora,* simplesmente, de forma rigorosa, já que, a partir de critérios estabelecidos, um *corpus* pode apresentar uma amostra de uma determinada língua.

Atualmente, há discussões acerca do lugar da LC, ou seja, trata-se de uma metodologia, uma teoria ou uma abordagem. Parodi (2008) traz autores como Stubbs (1996) e Tognini-Bonelli (2001) que defendem que a LC está se mostrando como uma teoria em face da associação com as tecnologias da informática. Novodvorski e Finatto (2014) afirmam que

> a expansão do uso dos termos *corpus* e *corpora*, além da menção a muitas das ferramentas e princípios caros à LC, alcança áreas que poderiam parecer, num primeiro momento, incompatíveis ou inimagináveis. Assim, a alusão às terminologias típicas de LC (como *types*, *tokens* e concordâncias) vem se tornando cada vez mais recorrente. Em eventos científicos, em publicações, em nomes de disciplinas, teses e dissertações, a recorrência com que aparecem referências ou vestígios da LC denotam já uma presença marcada no plano acadêmico e servem como um bom termômetro do estado da arte. (NOVODVORSKI; FINATTO, 2014, p. 8).

A LC conta, então, com princípios orientadores e originais e conta com o desenvolvimento de programas sofisticados e imprescindíveis para as pesquisas (PARODI, 2008). O autor ainda afirma que a LC constitui um conjunto de princípios metodológicos para estudar qualquer domínio linguístico e que se caracteriza por fornecer subsídios à investigação de uma língua em uso.

A LC pode atender a todos os ramos da linguística, pois é um método de investigação que pode ser aplicado em todos os níveis da língua e nos diferentes enfoques. Ratificando, Novodvorski e Finatto (2014, p. 8) afirmam que a "LC também é um modo de compreender a língua, que temos nosso modo de defini-la como objeto de estudo: a língua é um sistema probabilístico de combinatórias". Não se inscreve, portanto, em nenhuma filiação teórica, ou seja, não se pode afirmar que a LC seja uma teoria, mas ela serve a todos que desejam usufruir de sua metodologia. Scott (2010) afirma que a LC se constitui como uma abordagem

metodológica para o estudo de quaisquer línguas e pode ser considerada uma oportunidade revolucionária para a descrição, análise e até ensino de todos os tipos de discursos. Constitui, então, um conjunto de princípios metodológicos para estudar qualquer domínio linguístico e fornece subsídios para as pesquisas em língua em uso a partir de um *corpus* linguístico, desde que apoiado por tecnologias de informática e *software*. Para além disso, ainda conforme Parodi (2008), a LC fornece uma base empírica para a construção de dicionários, de gramáticas dentre outros.

A LC fornece subsídios para estudos em quaisquer ramos da Linguística e de outras áreas do saber, em que há trabalhos díspares como a Ciência Política, Agronomia, Jornalismo, Direito, Educação, etc. Novodvorski e Finatto (2014) chamam a atenção para o fato de que, embora nenhum *corpus* ofereça resposta para tudo aquilo que o pesquisador teria expectativa de identificar, numa perspectiva de pesquisa baseada em *corpus*, a abordagem de investigação direcionada pelo *corpus*, tal como a adotada nesta tese, revela aspectos que poderiam ser intuídos, antes do início dos trabalhos:

> todo *corpus* sempre traz questões novas ou questões que não se imaginava encontrar, ainda que, – de acordo com o próprio Fillmore (1992) – nenhum *corpus* nos dê resposta para tudo. De tal modo, tanto as observações como os experimentos e hipóteses formuladas no âmbito de toda investigação nos conduzem a uma revisão à luz das comprovações e dos resultados. (NOVODVORSKI; FINATTO, 2014, p. 9).

Isso leva à reflexão de que a partir da análise de um *corpus* e da sistematização dos dados, a hipótese inicial pode ser refutada ou confirmada. Os dados falam por si só, ou seja, o pesquisador, ao iniciar seu trabalho de investigação, não tem clareza ou total certeza do que encontrará, uma vez que, em determinado momento, as pesquisas passariam a ser guiadas pelo *corpus*. Assim, Novodvorski e Finatto (2014, p. 9), afirmam que "a sistematização de dados e de observações chega a ser crucial. Talvez ainda mais importante do que a simples aplicação e contraste de teorias". Em vista disso, a observação e a identificação de padrões são relevantes para o pesquisador, pois a partir da manipulação dos dados é que as análises serão conduzidas.

Sendo assim, a LC promove um estudo empírico, a partir da observação e descrição dos fatos linguísticos, por meio de ferramentas computacionais. A LC possui métodos que

garantem o estudo de evidências linguísticas sobre determinada língua ou sobre um fenômeno linguístico.

Para o tratamento de grandes quantidades de textos e abstração de palavras-chave recorremos à ferramenta *KeyWords* que gera uma lista de palavras consideradas chave para a pesquisa. Essa ferramenta provê o pesquisador de dados de caráter quantitativo com a possibilidade de escolha da maneira de observar esses dados que pode ser por ordem alfabética, ordem de chavicidade, dentre outros, considerando a adequação aos propósitos do pesquisador. A chavicidade explicita o quão importante cada palavra-chave é para o *corpus* de estudo, ou seja, ela indica a frequência dessa palavra dentro de cada *corpus*. Sendo assim, na seção seguinte, traçamos alguns apontamentos sobre Chavicidade.

3.9.2 Chavicidade

Uma investigação, por meio da chavicidade, pode levar às temáticas de um *corpus*, dentre outros aspectos. Segundo Novodvorski (2013, p. 55), com base em Scott (2010), "a análise da chavicidade (*Keyness*) está começando a despertar o interesse de pesquisadores, como uma qualidade textual que daria fortes indícios sobre a temática do texto, junto a indicadores de estilo". Isso porque a chavicidade é uma qualidade intrínseca de um texto, porém pode não ser de uma língua em si, isto é, a palavra pode ser chave em determinado texto, mas não ser em outro texto ou *corpus* da mesma língua, ou pode ser chave em um *corpus* específico e não ser numa determinada língua.

Scott (2010) ainda afirma que as palavras-chave servem como um direcionamento para o pesquisador. Nesse sentido, a chavicidade "indica áreas que valeria a pena investigar, uma vez que essas palavras se tornam proeminentes por alguma razão que deveriam ser analisadas". (NOVODVORSKI, 2013, p. 55).

Palavras-chave "são aquelas cujas frequências são diferentes, de modo estatisticamente significativo, das encontradas em um *corpus* de referência". (BERBER SARDINHA, 2009, p. 210). Não se deve considerar, contudo, que palavras-chave são palavras mais importantes, segundo o autor. São consideradas palavras-chave com base no critério estatístico-quantitativo. Há dois componentes principais para análise de palavras-chave, segundo Berber Sardinha (2004, p. 7; 2009, p. 211): i) *corpus* de estudo, "representado

por uma lista de frequência de palavras. O *corpus* de estudo é aquele que se pretende descrever"; ii) *corpus* de referência, também é conhecido como *corpus* de controle, e funciona como termo de comparação para a análise". O autor aponta que o *corpus* de referência deveria ser entre três e cinco vezes maior do que o *corpus* de estudo.

A comparação entre o *corpus* de estudo e o *corpus* de referência é realizada por meio de estatística, possibilitada por um programa de computador, de acordo com a necessidade do usuário. Uma das possibilidades é verificar quais palavras são mais frequentes no *corpus* de estudo em relação ao *corpus* de referência; contudo, ainda mais interessante para este tipo de estudo é verificar quais das palavras-chave extraídas pela ferramenta serão mais chave, isto é, apresentarão mais chavicidade, independentemente de serem as mais frequentes no *corpus* de estudo. Para estabelecimento da chavicidade, o programa estabelece um cálculo por meio do contraste entre as porcentagens derivadas da frequência dos itens em cada um dos *corpora* de estudo e de referência. O valor resultante indica o potencial estatístico de significância dos vocábulos. Outra das virtudes deste tipo de estudo é o contraste entre os itens que reportam à frequência zero no *corpus* de referência. Ainda que não haja uma frequência elevada no *corpus* de estudo, a simples ocorrência desses vocábulos assinala ser uma área pertinente de análise para o pesquisador.

As palavras-chave são um indicador para o pesquisador. Indica o que é mais recorrente e, portanto, o que seria mais importante pesquisar, já que, se são palavras mais recorrentes, significa também que, por alguma razão, deveriam ser analisadas sob algum aspecto (SCOTT, 2010). Além disso, as palavras-chave, ainda segundo o mesmo autor, poderiam oferecer indícios sobre a temática do texto. Assim, tal estudo promete ser revelador de aspectos pertinentes ao léxico indianista de José de Alencar.

Stubbs (2010 apud NOVODVORSKI, 2013, p. 54) propõe uma análise da expressão palavras-chave em três significados diferentes: "i) dos estudos culturais; ii) da análise comparativa e quantitativa de *corpus*, que identifica palavras estatisticamente proeminentes em textos ou coleções de textos; e iii) do trabalho em lexicogramática". No caso do léxico, a frequência é, sobremaneira, relevante, pois apresentará a recorrência e, em consequência, poder-se-á falar em norma linguística. Tal afirmação pode ser referendada em Biderman (1998, p. 162) quando afirma que "a norma linguística se baseia na frequência dos usos linguísticos. Assim, a norma linguística nada mais é do que a média dos usos frequentes das palavras que são aceitas pelas comunidades dos falantes".

No caso desta pesquisa, o *corpus* de estudo será composto pelas três obras consideradas indianistas de José de Alencar e um dos *corpora* de referência, as demais obras do mesmo autor. O *corpus* será mais bem detalhado no Capítulo 4 *Corpus* e Metodologia. A seguir, na próxima seção, apresentamos brevemente algumas características das ferramentas do WST que utilizamos em nossa pesquisa: *WordList*, *KeyWord* e *Concord*.

3.9.3 Ferramentas de análise lexical: WordList, KeyWords *e* Concord

Os trabalhos de pesquisa em LC requerem a utilização de programas computacionais para análises lexicais, como é o caso particular da presente pesquisa. O *WST* é um dos programas mais utilizados nas pesquisas, por propiciar uma interface considerada "amigável" com o pesquisador. As ferramentas possibilitam, em primeiro lugar, gerar listas de palavras e, a partir delas, extrair palavras-chave e/ou linhas de concordâncias como os itens de busca em contexto. Há outras funcionalidades e instrumentos disponíveis no programa, porém nos ateremos às três utilizadas para realização desta pesquisa: *WordList*, *KeyWords* e *Concord*. A Figura 1 apresenta uma imagem parcial da tela do computador com o *WST* e as três ferramentas utilizadas.

Figura 1: Detalhe da tela inicial do *WST* com os nomes das ferramentas

Fonte: *WordSmith Tools* 6.0 (SCOTT, 2012)

Há uma variedade de programas com os quais se analisa um *corpus*, porém optamos pelo *WST* em razão de o programa atender às necessidades desta pesquisa e pela facilidade em manipulá-lo. O *WST* foi criado em 1996 por Mike Scott, da Universidade de Liverpool, Reino Unido, e comercializado pela *Oxford University Press*. O *WST* possui um conjunto de ferramentas que proporcionam a manipulação de *corpora* para análises linguísticas diversas.

Como mencionado, utilizamos as ferramentas *WordList*, *KeyWords* e *Concord* descritas sucintamente, conforme Berber Sardinha (2009, p. 9):

> - *WordList:* produz listas de palavras contendo todas as palavras do arquivo ou arquivos selecionados, elencados em conjunto com suas frequências absolutas e percentuais. Também compara listas, criando listas de consistência, onde é informado em quantas listas cada palavra aparece.
>
> - *Concord*: realiza concordâncias, ou listagens de uma palavra específica (o 'nódulo', *node word ou search word*) juntamente com parte do texto onde ocorreu. Oferece também listas de colocados, isto é, palavras que ocorreram perto do nódulo.
>
> - *KeyWords*: extrai palavras de uma lista cujas frequências são estatisticamente diferentes (maiores ou menores) do que as frequências das mesmas palavras num outro corpus (de referência). Calcula também palavras-chave chave, que são chave em vários textos.

A *WordList* permite também que listas de palavras, presentes no *corpus*, sejam geradas em ordem alfabética ou por ordem decrescente de frequência, além de dados estatísticos. A combinação destas ferramentas permitirá a extração do léxico indianista, assim como a observância e transcrição dos contextos abonatórios, por meio das linhas de concordância dos vocábulos.

Finalizados os apontamentos sobre os aportes teóricos desta pesquisa, no capítulo seguinte, detalharemos os aspectos relacionados ao nosso *corpus* de estudo, assim como, de maneira pormenorizada, elencaremos os procedimentos metodológicos de nossa pesquisa.

CAPÍTULO 4 - *CORPUS* E METODOLOGIA

Neste capítulo, apresentamos, inicialmente, os dados relativos aos *corpora* de estudo e de referência e, em seguida, de maneira detalhada, os procedimentos metodológicos relacionados ao planejamento, à compilação, ao armazenamento e à análise do *corpus* para a realização desta pesquisa. São relatados, também, os procedimentos tanto para as escolhas tomadas durante a compilação e limpeza dos *corpora*, assim como os procedimentos adotados para a identificação, extração, análise e descrição do léxico indianista, até o preenchimento das fichas lexicográficas e proposta de verbetes. Para esta pesquisa, foi utilizado o programa *WordSmith Tools* 6.0 (SCOTT, 2012) e suas três ferramentas: *Concord*, *KeyWords* e *WordList*.

Nas seções seguintes, caracterizamos nosso *corpus* de estudo e o de referência, explicitando os procedimentos metodológicos em cada etapa.

4.1 Corpus

Esta seção está dividida em cinco subseções que tratam do *corpus* no que se refere aos apontamentos sobre realizar a pesquisa comparando a tríade indianista do autor José de Alencar com as suas demais obras não indianistas. Explicitamos detalhes sobre o *corpus* de estudo e os *corpora* de referência, seguindo de apontamentos sobre os perfis de cada *corpus*. Finalizamos traçando um comentário sobre os demais *corpora* de referência e o *corpus* de consulta, o *Corpus do Português* (DAVIES, 2016).

4.1.1 Corpus *de José de Alencar: o autor com ele mesmo*

Quando José de Alencar começou a escrever seus romances, esse gênero era recente na criação literária brasileira, ainda sujeita às imposições clássicas do Arcadismo. Por esse motivo, dentre outros, as críticas sobre seus romances não foram condescendentes. Alencar, porém, conforme Abreu (2011, p. 15-16), "adensa as discussões em pauta e explicita os temas reputados como imprescindíveis à formação da nacionalidade". Assim os romances não

tinham a função de mera distração que o gênero poderia propor, mas havia também, talvez principalmente, "o propósito de fornecer e fundamentar os emblemas distintivos do país".

A despeito das obras indianistas, Alencar buscou construir e difundir o nacionalismo literário e reafirmar a necessidade de separação entre a língua utilizada no Brasil e em Portugal e buscou erigir, como símbolos da nação livre dos domínios linguísticos dos europeus, os índios e a natureza, considerados os maiores e melhores representantes da nova nação.

Porém, a grandiosidade da obra de Alencar não se resume às indianistas. Ele fez um retrato do Brasil em sua produção, pois descreveu desde a vida da sociedade burguesa do Rio de Janeiro até o sertanejo de regiões mais impensadas e afastadas do país. Em razão disso, os romances são, normalmente, divididos em temas: urbanos, indianistas, regionalistas e históricos, além das cartas e das peças de teatro.[17]

Considerando a vasta produção de Alencar, cogitamos a possibilidade de estabelecer tanto o *corpus* de estudo quanto o de referência no conjunto da obra do próprio autor, ou seja, num primeiro momento, tivemos o interesse em observar a produtividade das palavras-chave, a partir do contraste de Alencar com ele mesmo. Para isso, as obras indianistas comporiam o *corpus* de estudo e as não indianistas o *corpus* de referência.

Tomando como base o que propõe Berber Sardinha (2004; 2009) de que o *corpus* de referência deve ser, no mínimo, cinco vezes maior em número de *itens*, que o *corpus* de estudo e, partindo da premissa inicial da pesquisa de que José de Alencar utiliza um léxico especializado nas obras indianistas, realizamos uma busca pelas obras do autor, a fim de se verificar se estudar o autor com ele mesmo atenderia a este critério de extensão. Após o levantamento das obras do autor disponíveis para acesso, produzimos uma lista de palavras com a qual foi constatado, em número de palavras, que o *corpus* de referência é mais do que cinco vezes maior que o de estudo, o que valida a análise.

A Tabela 2 apresenta um resumo do *corpus* de estudo em comparação com o *corpus* de referência. O *corpus* de estudo corresponde às três obras indianistas: *Iracema*, *O Guarani* e *Ubirajara* e o *corpus* de referência às demais obras do autor classificadas como não indianistas. Como é possível constatar, a partir do número de itens do *corpus* de estudo

[17] A descrição detalhada da produção de Alencar será apresentada na seção que trata do *corpus* de estudos e *corpus* de referência.

(150.524) em relação ao do *corpus* de referência (1.283.456), que o *corpus* de referência atende ao critério extensão em relação ao *corpus* de estudo, conforme sugerido por Berber Sardinha (2004, 2009).

Tabela 2: *Corpus* de estudo e *corpus* de referência: o autor com ele mesmo

Corpus	Nº de Textos	Itens (tokens)	Formas (types)
Corpus de estudo	3	150.524	14.766
Corpus de Referência	27	1.283.456	61.121

Fonte: A autora, com base na ferramenta *WST*

A partir desse contraste, decidimos estabelecer tanto o *corpus* de estudo quanto o de referência com os textos do próprio autor José de Alencar, pois um dos nossos propósitos de pesquisa de examinar o léxico indianista em contraste com as demais obras do autor poderia ser atendido.

No princípio da pesquisa, pensamos ser possível utilizar as edições mais antigas, da 1ª a 3ª edições dos livros de Alencar, porém ao analisar alguns trechos como os seguintes, percebemos a inviabilidade desta opção.

(a) "E si era um guerreiro que plantava, a aypim endurecia como o páo d'arco", da primeira edição de *Ubirajara*;
(b) "Dahi a alguns momentos retirava do buraco com um desses vasos vidrado, a que os índios chamavão camuci";
(c) "Os guerreiros tabajaras excitados com as copiosas libações do espumante cauim, se inflamam á voz de Irapuam que tantas vezes os guiou ao combate, quantas á victoria";
(d) "Poty scismava. Em sua cabeça de mancebo morava o espírito de um abaetê" e "Quando teu filho deixar o seio de Iracema, ella morrerá, como o abaty depois que deu seu fructo"[18];

[18] Os trechos citados compõem os verbetes do *Dicionário Histórico das Palavras Portuguesas de origem Tupi*, de Antônio Geraldo da Cunha de 1924.

É possível observar que grande parte das palavras como “scismava”, “ella”, “fructo”, “aypim”, “Irapuam” estão registradas com a grafia anterior às mudanças fonéticas e morfológicas propostas pelos acordos ortográficos. Esse fato inviabilizou a proposta deste estudo com base nas primeiras edições, primeiro, porque não encontramos todas as obras em suas primeiras edições disponíveis para acesso, para realizar o estudo contrastando com as obras de mesma grafia; segundo, porque os consulentes atuais, para os quais se destina o Vocabulário proposto, também não têm acesso às edições mais antigas. Definimos, portanto, que seriam utilizadas as obras editadas com atualizações posteriores ao Acordo Ortográfico de 1990.

Importante salientar que apresentamos os anos de publicação das primeiras edições de cada obra que compõem o *corpus* de estudo e o *corpus* de referências, porém todas as obras que compõem os *corpora* deste estudo são edições posteriores ao Acordo Ortográfico, entre os países cuja língua oficial é a Língua Portuguesa, firmado em 16 de dezembro de 1990[19].

Na seção seguinte, serão expostos mais detalhes sobre o *corpus* de estudo e o *corpus* de referência.

4.1.2 Corpus *de estudo e* corpora *de referência*

Os textos de José de Alencar que compõem o *corpus* de estudo e o de referência estão disponíveis no site www.dominiopublico.gov.br[20], exceto três obras do *corpus* de referências: *O Garatuja*[21], *O Ermitão da Glória*[22] *e Cartas*[23]. Não foi possível localizar a *Carta sobre A Confederação dos Tamoios* (1856); as Cartas *O juízo de Deus* (1867) e *Visão de Jó* (1867); e a pela de teatro *O jesuíta* (1857) em PDF, para acesso. Entretanto, a ausência destes textos

[19] Dados a partir do Portal do MEC. Disponível em: <http://portal.mec.gov.br/arquivos/pdf/acordoortografico.pdf>. Acesso em: 27 jul. 2017.
20 A consulta ao site foi realizada no dia 27 de jul. de 2017.
21 Disponível em: <http://www.rio.rj.gov.br/dlstatic/10112/4204210/4101376/garatuja.pdf>. Acesso em: 27 jul. 2017.
22 Disponível em: <http://www.portugues.seed.pr.gov.br/arquivos/File/leit_online/jose17.pdf>. Acesso em: 27 jul. 2017.
23 Disponível em: <http://www.academia.org.br/sites/default/files/publicacoes/arquivos/cartas_de_erasmo_ao_imperador_-_jose_de_alencar.pdf>. Acesso em: 27 jul. 2017.

não invalida a pesquisa, tendo em vista que o *corpus* de referência atende às exigências da pesquisa, como já mencionado.

Um dos objetivos desta pesquisa, como já apontado, é analisar o léxico das obras indianistas de Alencar em contraste com as demais obras do mesmo autor. Ainda assim, buscamos também contrastar os resultados a partir de outros três *corpora* de referência, a saber: CorpRef-Lácio-Web, CorpRef-AcadTeses, e o CorpRef-Nov. Essa testagem busca verificar a produtividade de Alencar em dois aspectos: primeiro, contrastar Alencar com ele mesmo, por meio do contraste do *corpus* de estudo com o *corpus* de referência também de Alencar, CorpRef-Alencar e, ao mesmo tempo, poder comparar esses resultados com os resultados utilizando outros *corpora* de extensões e características diferentes. Com esse procedimento, verificaremos os resultados com vistas a analisar as palavras-chave do *corpus* de estudo com os demais *corpora*. Procedemos, a seguir, a uma caracterização dos *corpora* de referência.

O CorpRef-Lácio-Web é um projeto iniciado no início de 2002, com parceria entre NILC (Núcleo Interinstitucional de Linguística Computacional) e está localizado no ICMC-USP, IME (Instituto de Matemática e Estatística) e FFLCH (Faculdade de Filosofia, Letras e Ciências Humanas). "O objetivo deste projeto é disponibilizar na *Web*, vários *corpora* do português brasileiro contemporâneo, representando bancos de textos adequadamente compilados, catalogados e codificados em um padrão que possibilite fácil intercâmbio, navegação e análise"; assim como "ferramentas linguístico-computacionais, tais como contadores de frequência, concordanciadores e etiquetadores morfossintáticos". O público a que se destina o Lácio *Web* é heterogêneo: de um lado linguistas, cientistas da computação, lexicógrafos, da mesma forma a não especialistas em geral, já que os textos estão separados por áreas de conhecimento[24].

CorpRef-AcadTeses é um *corpus* que foi compilado no âmbito de um outro projeto, o do *Corpus* Brasileiro, de Berber Sardinha. A lista de palavras desse *corpus* foi disponibilizada por contato com o orientador deste trabalho com o pesquisador[25]. Trata-se de um *corpus* acadêmico formado exclusivamente por textos escritos a partir de um banco de teses. Ressaltamos que esse título foi escolhido por nós (orientador e eu), já que utilizamos somente a parte escrita acadêmica de teses do *Corpus* Brasileiro.

[24] Dados obtidos no sítio: < http://143.107.183.175:22180/lacioweb/descricao.htm>. Acesso em 12 ago. 2017.
[25] Disponível em: <http://corpusbrasileiro.pucsp.br/cb/Inicial.html> Acesso em: 09 maio 2018. Ressaltamos que tivemos acesso apenas à Lista de palavras do segmento acadêmico, ou seja, não tivemos acesso ao *corpus*.

O CorpRef-Nov foi compilado por Novodvorski (2013) e disponibilizado para uso desta pesquisa. A compilação ocorreu durante a sua pesquisa de Doutorado, quando compilou textos jornalísticos, acadêmicos e literários, via *internet*, em proporções contrabalanceadas, que guardassem o equilíbrio em termos quantitativos de extensão.

A Tabela 3 apresenta as características do *corpus* de estudo em comparação aos *corpora* de referência, obtidos por meio da ferramenta *WordList* do *WST*.

Tabela 3: Comparação entre o *corpus* de estudo e os *corpora* de referência

Corpus	Nº de Textos	Itens (tokens)	Formas (types)	Razão Forma/Item	Razão Forma/Item padronizada
CorpEST	3	150.524	14.766	9.81	47,50
CorpRef-Alencar	27	1.283.456	61.121	7,76	50,61
CorpRef-Lácio-Web	6.240	7.054.763	130.020	1,84	43,26
CorpRef-AcadTeses	11.392	96.669.768	620.068	0,64	40,51
CorpRef-Nov	211	500.523	43.119	8,61	48,69

Fonte: A autora, com base na ferramenta *WST*

É importante dimensionar a frequência com que determinados itens lexicais são empregados por Alencar, para averiguarmos se o autor utiliza um léxico especial indianista em seus romances que compõem a trilogia indianista do autor.

4.1.3 Perfil dos textos que compõem o corpus *de estudo*

A composição do *corpus* linguístico de estudo, que será analisado neste trabalho, reúne as três obras consideradas indianistas do autor José de Alencar. Segundo Berber Sardinha (2009, p. 194) "um *corpus* de estudo, representado em uma lista de frequência de palavras. O *corpus* de estudo é aquele que se pretende descrever. A ferramenta *KeyWords* aceita a análise simultânea de mais de um *corpus* de estudo". O Quadro 4 apresenta as três obras de Alencar que compõem o *corpus* de estudo, acompanhadas com a indicação do ano de publicação da primeira edição.

Quadro 4: *Corpus* de estudo

Obra	Ano da 1ª edição
O Guarani	1857
Iracema	1865
Ubirajara	1874

Fonte: A autora

Para abordar o léxico indianista de Alencar, inicialmente, optamos pelo programa *WordSmith Tools*, de Mike Scott, na versão 6.0 de *WST* (SCOTT, 2012), com a função *Statistics*, que nos apresentou os dados da Figura 2: *corpus* de estudo. Embora não pretendamos repetir um manual para utilização das ferramentas do programa *WordSmith Tools,* é conveniente recorrer a Berber Sardinha (2009, p. 161) e apresentar, sucintamente, as nomenclaturas utilizadas.

- *Text File*. Mostra o nome de cada arquivo processado e a palavra "*Overall*" é usada para indicar o total dos arquivos.

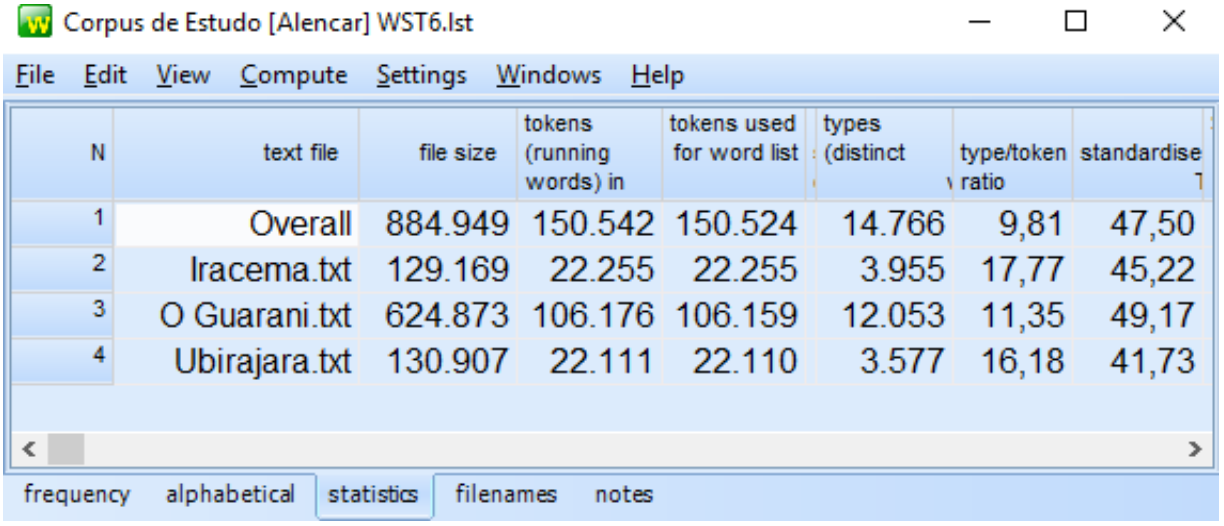
- *Tokens*. Indica o total de ocorrências de palavras no texto. Significa o total de palavras, levando em conta as repetições, desde a primeira até a última de todos os arquivos selecionados.

- *Types*. Mostra o total de itens sem levar em conta as repetições.

- *Type-Token Ratio.* TTR - Resultado da divisão do total de *Types* pelo total de *Tokens*, multiplicado por 100. Obtém, com essa operação, a extensão da variação lexical do texto.

A Figura 2 apresenta os dados sobre o c*orpus* de estudo obtidos a partir da ferramenta *WordList*, disponibilizada pelo programa, a partir do ícone S*tatistics*.

Figura 2: *Corpus* de estudo



Fonte: A autora, por meio da Ferramenta *Statistics* do *WST* 6.0.

É possível observar os dados estatísticos referentes ao *corpus* de estudo. O texto com maior número de palavras é *O Guarani* e o menor, considerando o mesmo aspecto é *Ubirajara*. A seguir descrevemos o *corpus* de referência de Alencar, o CorpRef-Alencar.

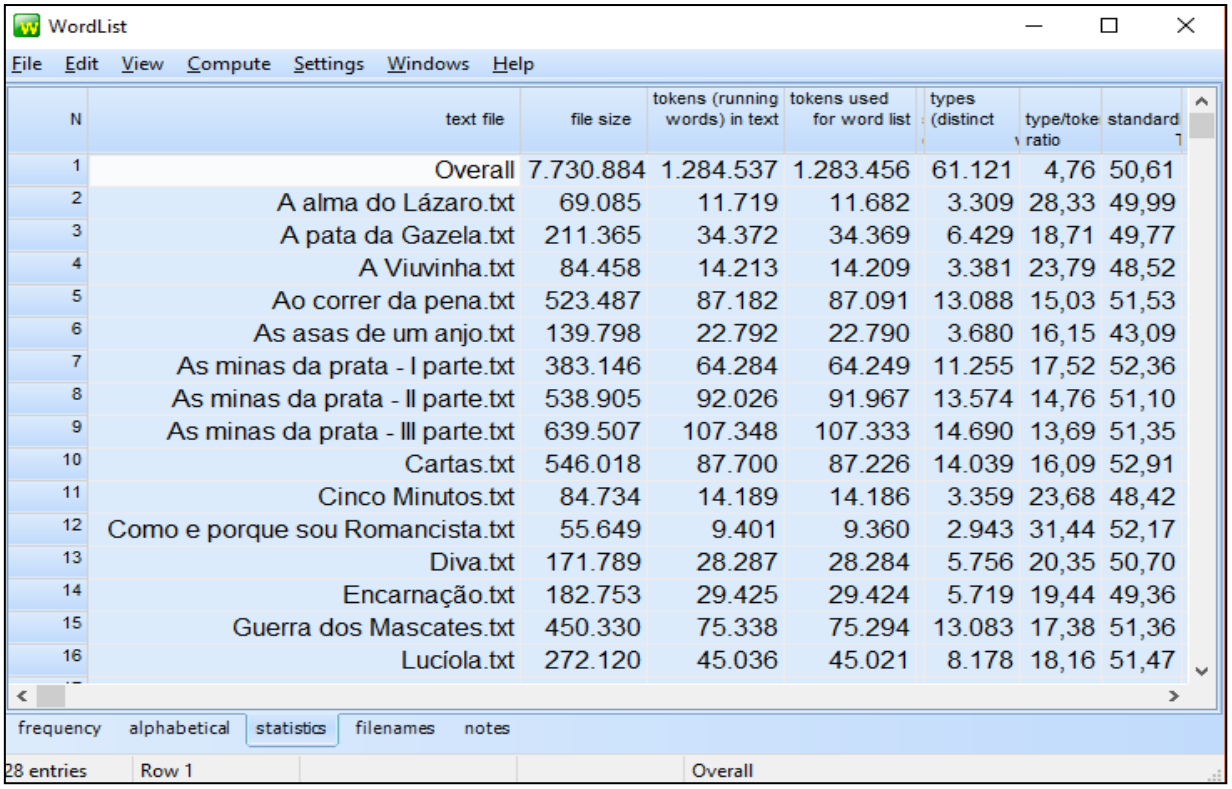
4.1.4 Perfil dos textos de José de Alencar que compõem o corpus *de referência*

Como mencionado, a fim de atender aos objetivos propostos nesta tese, um dos *corpora* de referência deste estudo também é formado por obras do mesmo autor do *corpus* de estudo, visto que se pretende comparar o autor com ele mesmo. Para Berber Sardinha (1999, p. 15) "o tamanho do *corpus* de referência é um dos cinco elementos que podem influenciar o resultado de uma análise por palavras-chave, no tocante à quantidade de palavras-chave que podem ser obtidas". O *corpus* de referência, Segundo Berber Sardinha (2009, p. 194), "também é conhecido como *corpus* de controle, e funciona como termo de comparação para a análise. A sua função é a de fornecer uma norma com a qual se fará a comparação das frequências do *corpus* de estudo". Esta comparação se dá por meio de uma prova estatística selecionada pelo usuário.

Com o propósito de reforçar a justificativa da possibilidade de se contrastar o autor com ele mesmo, apresentamos uma vista parcial do *corpus* de referência. Como não é possível capturar da tela do *WST* a sua totalidade, apresentamos uma vista parcial dos 16

primeiros títulos listados pelo programa, assim como os números referentes aos Itens, Formas, Razão Forma/Item, Razão Forma/Item padronizada, conforme apresentado na Figura 3.

Figura 3: *Corpus* de Referência



N	text file	file size	tokens (running words) in text	tokens used for word list	types (distinct	type/toke ratio	standard T
1	Overall	7.730.884	1.284.537	1.283.456	61.121	4,76	50,61
2	A alma do Lázaro.txt	69.085	11.719	11.682	3.309	28,33	49,99
3	A pata da Gazela.txt	211.365	34.372	34.369	6.429	18,71	49,77
4	A Viuvinha.txt	84.458	14.213	14.209	3.381	23,79	48,52
5	Ao correr da pena.txt	523.487	87.182	87.091	13.088	15,03	51,53
6	As asas de um anjo.txt	139.798	22.792	22.790	3.680	16,15	43,09
7	As minas da prata - I parte.txt	383.146	64.284	64.249	11.255	17,52	52,36
8	As minas da prata - II parte.txt	538.905	92.026	91.967	13.574	14,76	51,10
9	As minas da prata - III parte.txt	639.507	107.348	107.333	14.690	13,69	51,35
10	Cartas.txt	546.018	87.700	87.226	14.039	16,09	52,91
11	Cinco Minutos.txt	84.734	14.189	14.186	3.359	23,68	48,42
12	Como e porque sou Romancista.txt	55.649	9.401	9.360	2.943	31,44	52,17
13	Diva.txt	171.789	28.287	28.284	5.756	20,35	50,70
14	Encarnação.txt	182.753	29.425	29.424	5.719	19,44	49,36
15	Guerra dos Mascates.txt	450.330	75.338	75.294	13.083	17,38	51,36
16	Lucíola.txt	272.120	45.036	45.021	8.178	18,16	51,47

Fonte: A autora, por meio da Ferramenta *WordList* do *WST*

Como não é possível listar os títulos de todos os textos de Alencar que compõem o *corpus* de referência, apresentamo-los no Quadro 5 em ordem alfabética, seguida pelo ano da primeira edição e a classificação das obras em romance, teatro, carta ou autobiografia. Vale ressaltar que as edições utilizadas nesta pesquisa são posteriores ao ano de 1990, ano da Reforma ortográfica da Língua Portuguesa. Não consideramos relevante o ano exato da edição utilizada no contraste desta pesquisa, portanto, essa datação não será mencionada.

Quadro 5: *Corpus* de Referência CorpRef-Alencar

	Obra	Ano da 1ª edição	Classificação
1	A alma do Lázaro	1872	Romance Histórico

2	A Pata da Gazela	1870	Romance Urbano
3	A Viuvinha	1857	Romance Urbano
4	Ao correr da pena	1854-55	Romance Crônicas
5	A asas de um anjo	1858	Teatro
6	As Minas da Prata – I parte	1864-65	Romance Histórico
7	As Minas da Prata – II parte	1864-65	Romance Histórico
8	As Minas da Prata – III parte	1864-65	Romance Histórico
9	Cartas[26]	1865-68	Cartas
10	Cinco Minutos	1856	Romance Urbano
11	Como e Porque sou Romancista	1893	Autobiografia
12	Diva	1864	Romance Urbano
13	Encarnação	1877	Romance Urbano
14	Guerra dos Mascates	1873	Romance Histórico
15	Lucíola	1862	Romance Urbano
16	Mãe	1862	Teatro
17	O demônio familiar	1857	Teatro
18	O Ermitão da Glória	1873	Romance Histórico
19	O Garatuja	1873	Romance Histórico
20	O Gaúcho	1870	Romance Regionalista
21	O que é o Casamento?	1861	Teatro
22	O Rio de Janeiro - Verso e Reverso	1857	Teatro
23	O Sertanejo	1875	Romance Regionalista
24	O tronco do Ipê	1871	Romance Regionalista
25	Senhora	1875	Romance Urbano

[26] Neste volume estão incluídas 5 cartas: Ao Imperador, Cartas de Erasmo; Ao povo: cartas políticas de Erasmo; Ao Visconde de Itaboraí, Carta de Erasmo sobre a Crise Financeira; Ao Marquês de Olinda; Ao Imperador, Novas Cartas Políticas de Erasmo

26	Sonhos d'Oro	1872	Romance Urbano
27	Til	1872	Romance Regionalista

Fonte: A autora

Após a apresentação do *corpus* de estudo e *corpus* de referência de Alencar, é momento de informar sobre o *corpus* de consulta.

4.1.5 Corpus *de consulta*

Inventariar todas as unidades lexicais de uma língua é uma tarefa impossível, considerando que uma pessoa utiliza, ao longo de sua vida, uma imensurável quantidade de vocábulos e, se somarmos esse número com os demais membros de uma mesma língua e, levando em conta também o aspecto diacrônico, chegaríamos a uma quantidade astronômica de vocábulos. Sendo assim, mesmo que tratemos do léxico de uma língua, em particular, não seria possível abarcar todas as unidades lexicais dessa língua. Viaro (2014) estabelece uma comparação entre a Linguística e a Biologia. Esta última, se preocupa com os seres vivos e estuda as espécies, não cada indivíduo dessa espécie; assim é com a Linguística, cujo objeto de estudo são as abstrações, a partir dos estudos dos fatos, já que a totalidade dos dados produzidos por um indivíduo sozinho, não podem ser tomados como um comportamento geral.

Em razão disso, o tratamento dos *corpora*, nos estudos linguísticos atuais, tem recebido uma atenção especial (Sardinha, 2004). Segundo Viaro (2014, p. 101) "nos estudos etimológicos, os *corpora* são de grande importância, sobretudo os organizados de forma que se possa obter alguma informação diacrônica". O autor ratifica que os *corpora* eletrônicos são bastante úteis, quando se trata de questões históricas e sugere alguns *corpora* para pesquisa: o *Corpus do Português*; o *Corpus histórico do Português Tycho Brahe*; o *Corpus informatizado do Português Medieval*; o *Corpus de referência do Português Contemporâneo*; a *Linguateca* e o *Corpus Lexicográfico do Português*.

O autor ressalta que *todo corpus* que lide com textos produzidos antes da invenção da imprensa necessita de uma pessoa que domine Ecdótica[27]. Assim é necessário que um etimólogo consulte outros documentos, como as críticas às edições, à procura de alguma variante não mencionada pelo autor ou pelos editores. Nesse sentido, adotamos, nesta pesquisa, diversas fontes de investigação dos vocábulos, a fim de evitar o quanto possível o viés de um único autor. Assim sendo, utilizamos como fonte de pesquisa o *Corpus do Português* (Davies, 2016), o primeiro *corpus* sugerido por Viaro (2014) para consulta.

O *Corpus do Português*[28] (Davies, 2016) possibilita esquadrinhar a língua portuguesa e pesquisar a popularidade de palavras ou de frases entre os milhares de textos. É um *corpus* linguístico de textos da Língua Portuguesa e foi compilado pelos pesquisadores Mark Davies da Universidade Brigham Younf e Michael J. Ferreira da Universidade de Georgetown.

A interface permite pesquisas por palavras ou por frases, assim como por associações de palavras dentro de uma distância de até 10 palavras. Permite também a comparação entre a frequência e a distribuição de palavras, frases e construções gramaticais entre textos, com base em três possibilidades: i) por registro que são comparações entre os textos de origens diversas como texto coloquial, ficcional, jornalístico e acadêmico; ii) por dialeto que se pode estabelecer comparações entre os textos dos países com Brasil e Portugal; iii) por período histórico que trata da possibilidade de busca com datação, desde o século XIV ao XX.

Como descrito no *site*, o *Corpus do Português* é dividido em duas partes: um *corpus* considerado como original e menor e que permite ver as mudanças históricas e as variações de gênero; e um outro, denominado novo e maior que permite verificar as variações entre o português brasileiro e o português europeu na contemporaneidade.

De acordo com a descrição disponível, a versão histórica do *corpus* contém em torno de 45 milhões de palavras de aproximadamente 57.000 textos em português entre os anos de 1300 a 1900. Os textos compilados para organização deste *corpus* abrangem diversos gêneros como de conversação, ficção, jornais e acadêmicos. Assim sendo, o *corpus* permite comparar diversos gêneros e diversos períodos de tempo. A versão de 2016, segundo dados do *site*, contém um bilhão de palavras de textos em Português, abrangendo quatro países falantes da

[27] Segundo o dicionário Houaiss Ecdótica é ciência que busca, por meio de minuciosas regras de hermenêutica e exegese, restituir a forma mais próxima do que seria a redação inicial de um texto, a fim de que se estabeleça a sua edição definitiva; crítica textual.
[28] Disponível em: www.corpusdoportugues.org. Acesso em: 09 maio 2018.

Língua Portuguesa. O *Corpus do Português* permite também comparar a frequência de palavras, frases e construções gramaticais entre os quatro países.

Para esta pesquisa, utilizamos a aba **Gênero/Histórico**, o que nos permite verificar os empregos dos vocábulos de Alencar, identificando o período cuja datação é realizada por meio da indicação dos séculos. Essa consulta nos permite analisar, por meio de comparações com os dicionários de exclusão, o processo criativo do autor. A título de exemplo, a Figura 4 apresenta a palavra **anajê**, cuja primeira datação é a obra *Iracema* de Alencar.

Figura 4: Dados referentes a palavra **anajê** no *Corpus do Português*

Fonte: A autora, a partir do *Corpus do Português* (DAVIES, 2016)

Após a breve explicitação sobre o *corpus* de consulta, relatamos detalhadamente os procedimentos metodológicos.

4.2 Procedimentos Metodológicos

Nesta seção, são enumerados os procedimentos metodológicos aplicados desde o levantamento, compilação e tratamento do *corpus* até a descrição e análise dos dados. Vale ressaltar que, à medida que o estudo foi progredindo, novos procedimentos metodológicos

foram exigidos. À medida que a observação e análise do *corpus* ocorria, novas possibilidades de pesquisa foram surgindo. Em razão disso, decidimos adotar a perspectiva de análise guiada pelo *corpus*, conforme Berber Sardinha (2004, 2009), que nos revelou aspectos profícuos para análises. A abordagem direcionada pelo *corpus* nos permite entender que um *corpus* não é apenas um repositório utilizado apenas para validar uma teoria pré-existente ou pré-definida, conforme Tognini-Bonelli (2001)

> A teoria não tem existência independente das evidências e o percurso metodológico geral é claro: observação leva às hipóteses que levam a generalizações que levam à unificação em um construto teórico. É importante entender que essa metodologia [direcionada por *corpus*] não é mecânica, mas constantemente mediada pelo linguista, que ainda está se comportando como um linguista e aplicando seu conhecimento, experiência e inteligência em cada etapa durante o processo. Não existe indução pura. (TOGNINI-BONELLI, 2001, apud CARNEIRO, 2016, p. 84-85).

Considerando, portanto, o centro de interesse de nossa pesquisa e aplicando os nossos conhecimentos sobre Alencar no processo de mediação entre pesquisadora e *corpus*, esta pesquisa foi executada percorrendo os seguintes passos:

1) Levantamento e compilação dos *corpora* de estudo e de referência de José de Alencar
2) Armazenamento dos *corpora*;
3) Preparação do c*orpus* de estudo*;*
4) Extração das listas de palavras para o levantamento dos dados estatísticos;
5) Extração das listas de palavras-chave;
6) Análise contrastiva das diferentes listas de palavras-chave, a partir dos diferentes *corpora* de referência;
7) Identificação do léxico indianista;
8) Agrupamento do léxico indianista por campos semânticos;
9) Extração e análise das linhas de concordâncias com o léxico indianista;
10) Utilização de dicionários e de *corpus* de consulta, na análise e descrição etimológica de léxico de Alencar;
11) Elaboração das fichas lexicográficas com o léxico indianista identificado e analisado
12) Elaboração de proposta de verbetes;

4.2.1 Levantamento e compilação do **corpus** *de estudo e* **corpus** *de referência de José de Alencar*

Não é novidade afirmar que os trabalhos que utilizam a LC para o tratamento eletrônico de textos estão adquirindo notoriedade e avançando em quantidade e qualidade nos meios acadêmicos. Especificamente, trabalhos que utilizam a LC para fins de estudo de textos literários ainda são parcos, porém não há que se negar que a LC pode ser um suporte metodológico que colabora com os estudos, quando se pretende uma investigação que se utilize de dados estatísticos para as análises.

Para o levantamento do *corpus* desta pesquisa, realizamos, inicialmente, uma busca em *sites* em que se poderiam obter as obras de José de Alencar no formato PDF, cuja utilização estivesse livre de impedimentos ou restrições. Para a coleta do *corpus* desta pesquisa, então, seguimos os seguintes parâmetros de compilação: i) textos disponíveis no formato digital em PDF; ii) possibilidade de acesso sem custos ou restrições; iii) textos com a grafia adaptada tendo como base o acordo ortográfico de 1990. Na ferramenta *Google*, com as palavras "José de Alencar:pdf", encontramos vários textos advindos de diversos sítios, especialmente de universidades. Capturamos alguns textos e os analisamos, manualmente, comparando-os com as obras impressas que possuímos, para atestar a autenticidade e qualidade dos textos encontrados digitalizados nos diversos sítios. Ressaltamos que, para averiguar a autenticidade, comparamos apenas as obras digitalizadas do *corpus* de estudo: *Iracema*, *Ubirajara* e O *Guarani*[29] com as impressas que possuímos. As obras que compõem o *corpus* de referência não foram comparadas manualmente, pois não possuímos os volumes impressos.

Algumas obras foram descartadas, na primeira análise, pois não atendiam aos critérios acima mencionados. As obras disponibilizadas no Portal do Domínio Público[30] garantiram a fidedignidade e, portanto, desse *site*, capturamos todas as obras do *corpus* de estudo e a maioria das que compõem o *corpus* de referência. Excetuam-se, do *corpus* de referência, apenas *O Garatuja*, *O Ermitão da Glória* e *Cartas*, localizados em outros *sites* da internet, assim como há também aquelas que não foram localizadas como a *Carta sobre A Confederação dos Tamoios* (1856); as cartas *O juízo de Deus* (1867) e *Visão de Jó* (1867) e o teatro *O jesuíta* (1857), conforme já explicitado.

[29] Publicadas pela Livraria José Olympio.
[30] Disponível em: <www.dominiopublico.gov.br>. Acesso em 10 abr. 2018.

Como são textos com datação com mais de um século de publicação da primeira edição, as obras já são consideradas de domínio público, assim não houve a preocupação em obter autorização para uso dos textos, por não serem protegidos por direitos autorais. Não houve, também, a preocupação em preservar em sigilo os dados obtidos das obras, podendo, inclusive serem utilizados por outros pesquisadores. Os *corpora* de estudo e de referência de José de Alencar foram compilados e armazenados.

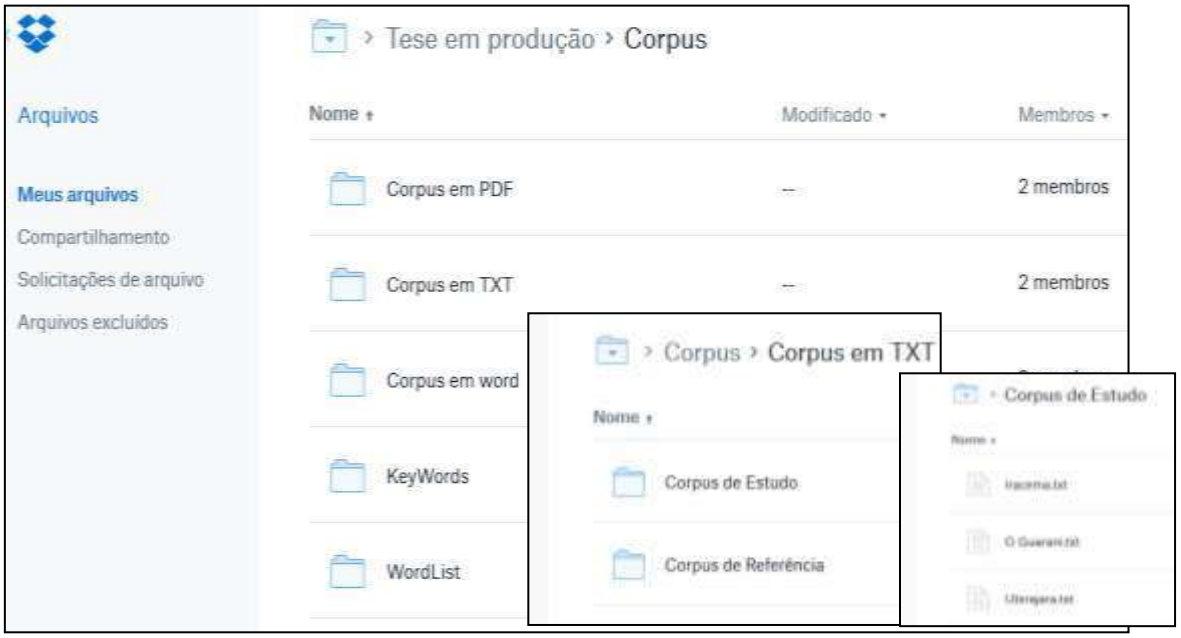
Na seção seguinte, detalhamos os procedimentos para o armazenamento dos *corpora*.

4.2.2 Armazenamento dos corpora

Após a compilação, os *corpora* foram armazenados em uma pasta denominada "*Corpus"*. Distribuímos os *corpora,* na pasta *Corpus,* em três subpastas: "*Corpus* em PDF", "*Corpus* em Word" e "*Corpus* em TXT". Os textos que não sofreram intervenção podem ser chamados de "*corpus* cru" e estão armazenados na subpasta "*Corpus* em PDF", subdivida em *Corpus* de estudo e *Corpus* de referência. Os *corpora* foram organizados dessa maneira com o propósito de manter os arquivos originais. Caso fosse necessário recorrer a eles, não haveria o trabalho de compilação, novamente.

O procedimento de armazenamento dos *corpora* em PDF, Word e TXT foi o mesmo, ou seja, nas pastas denominadas "*Corpus* em PDF", "*Corpus* em TXT" e "*Corpus* em Word", há duas subpastas com os "*Corpus* de Estudo" e "*Corpus* de Referência". Dentro das pastas "*Corpus* de Estudo" encontram-se as três obras que compõem o *corpus* de estudo e dentro as pastas "*Corpus* de Referência" os textos que compõem o *corpus* de referência de José de Alencar. Na Figura 5, observa-se uma imagem parcial do armazenamento dos *corpora* em TXT em pastas e subpastas.

Figura 5: Vista parcial do armazenamento dos *Corpora* de José de Alencar



Fonte: A autora, com base nas telas do *Dropbox*

Os *corpora* foram armazenados em um computador particular e, concomitantemente, na plataforma *Dropbox*, conforme demonstrado na Figura 5. Segundo Drago et al. (2013) os usuários do *Dropbox* podem sincronizar os arquivos por meio de uma mesma conta, podendo, inclusive, controlar os recursos de rede utilizados por cada membro. Assim, todas as pastas estão compartilhadas com o orientador, para que o trabalho entre pesquisadora e orientador possa ser melhor estruturado e facilitado também via *internet*.

Na seção seguinte, explicitaremos os procedimentos de preparação do *corpus* de estudo.

4.2.3 Preparação do corpus *de estudo*

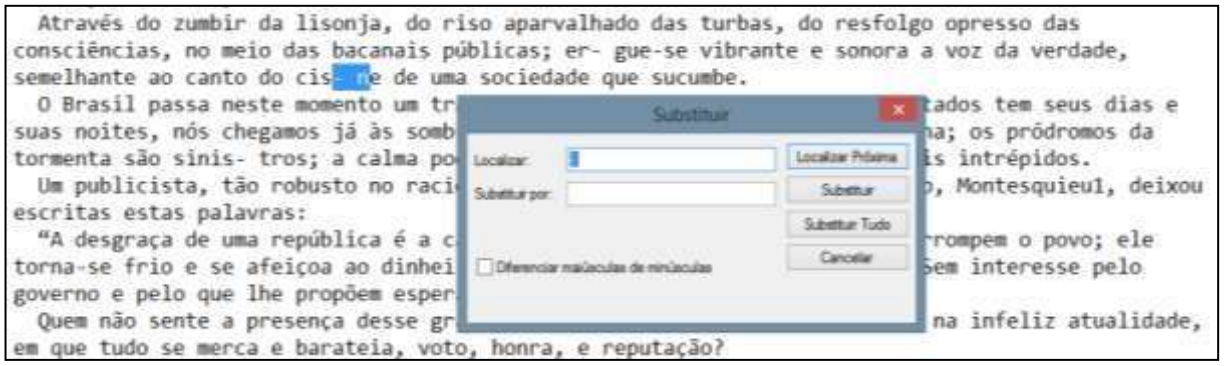
Assim que as obras em PDF foram compiladas e salvas em um diretório de um computador particular e na plataforma *Dropbox*[31], iniciamos a preparação do *corpus* para processamento pelo Programa *WST* em sua versão 6.0. Para isso, foi necessária a conversão de todos dos textos para o formato TXT (arquivos sem formatação). A preparação dos *corpora* antecede o processamento pelas ferramentas computacionais e compreende um

[31] O Dropbox é um serviço de armazenamento em nuvem online e possibilita o compartilhamento dos arquivos com outros usuários e permite acessá-los em qualquer lugar onde se possa acessar à internet.

procedimento metodológico. Nesta pesquisa, para a preparação dos *corpora*, adotamos dois procedimentos como testagens: a primeira, a conversão do formato em PDF para TXT, automaticamente; a segunda, testamos a conversão para a versão W*ord*[32], para, em seguida, a conversão para TXT.

A conversão direta dos textos em PDF para o TXT, não foi a mais viável. Nesse procedimento, surgiram alguns problemas ou erros que despenderiam um trabalho mais demorado e grandioso para a revisão, antes do início do processamento das próximas etapas da pesquisa. As Figuras 6 e 7 ilustram duas imagens parciais da tela do computador com o texto em formato TXT e dois dos procedimentos de limpeza e manejo do *corpus*. Na figura 6, foi realizado o processo de substituição dos cortes de palavras em razão da translineação como "cis- ne". Nesse caso substituímos o hífen e o espaço, como destacado em "cis- ne" por nada, ou seja, seria excluído

Figura 6: Procedimento de normalização do *corpus*

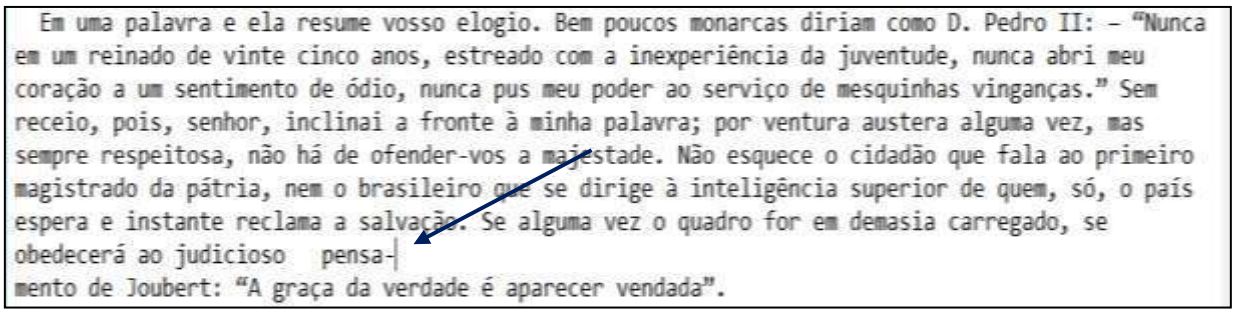


Fonte: A autora, extraído da tela do TXT

Com esse procedimento, os problemas de translineação ou quebra de palavras, na mesma linha do texto e que mantinham esse padrão, foram resolvidos. Porém ainda permaneceram outros, como os casos de translineação em que a segunda parte da palavra foi disposta na linha seguinte, como demonstra a Figura 7.

[32] A transposição do formato PDF para Word foi realizada com o auxílio da ferramenta disponibilizada no site https://smallpdf.com que converteu gratuitamente textos em PDF para Word. O site possui outras ferramentas de conversão para outros formatos de textos, porém não foram utilizadas.

Figura 7: Procedimento de revisão do texto



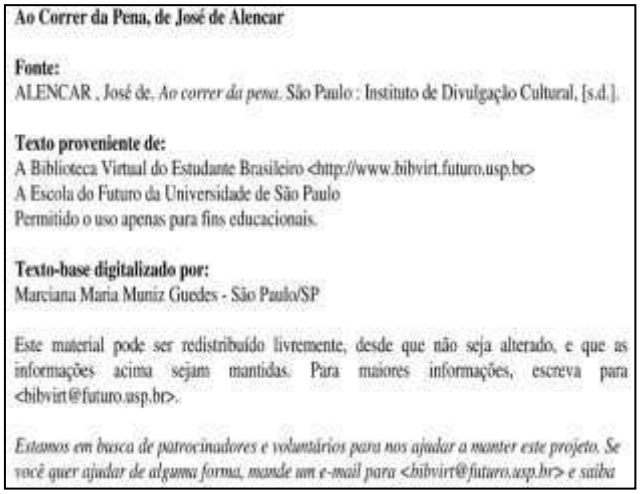
Fonte: A autora, extraído da tela do *TXT*

Nesse caso, demandaria uma revisão exaustiva de todo o texto, comparando-o com o original em PDF, já que o processo de substituição, como demonstrado na Figura 7, não cobriria. Além do exposto, outros problemas foram identificados com o texto já convertido no formato TXT. No momento da revisão, não foi possível identificar, por exemplo, o que se tratava de notas ou parte do corpo do texto, portanto seria necessário recorrer ao material em PDF para realizar a identificação e separação das notas do corpo do texto.

Como a conversão de PDF para TXT, diretamente, não foi a mais viável, testamos a segunda opção que foi a conversão, primeiramente, para *Word* e, depois de limpos e organizados, efetuar a conversão para TXT. Após a avaliação desses dois procedimentos, optou-se pela limpeza de todos os textos que compõem os *corpora* no formato *Word.* Depois da testagem e decidido o procedimento mais viável, procedemos à conversão de todos os textos que compõem o *corpus* de estudo e o de referência, para *Word*, para, assim, iniciarmos o trabalho de limpeza e revisão.

A limpeza do *corpus* compreende uma das etapas de preparação do *corpus* para processamento pelo *WST*. Outros procedimentos como a forma de armazenamento também são essenciais para o processamento dos *corpora* pelo *WST*, assim tivemos o cuidado de armazenar os *corpora* em pastas com rápido e fácil acesso. A limpeza tem como objetivo principal manter o conteúdo textual de cada obra retirando os elementos que não são processados pelo *WST* ou outras ferramentas computacionais ou não são de interesse da pesquisa, como, por exemplo, nome do autor, erros ortográficos, números de páginas ou de capítulos, dentre outros. As Figuras 8 e 9 demonstram um recorte de um texto em *Word* antes de iniciar os procedimentos de limpeza do *corpus*. Os dados apresentados nas Figura 8 são constantes do livro *Ao correr da pena* e os da Figura 9 do livro *O Ermitão da Glória.*

Figura 8: Elementos a serem removidos dos *corpora* de estudo e referência



Ao Correr da Pena, de José de Alencar

Fonte:
ALENCAR , José de. *Ao correr da pena*. São Paulo : Instituto de Divulgação Cultural, [s.d.].

Texto proveniente de:
A Biblioteca Virtual do Estudante Brasileiro <http://www.bibvirt.futuro.usp.br>
A Escola do Futuro da Universidade de São Paulo
Permitido o uso apenas para fins educacionais.

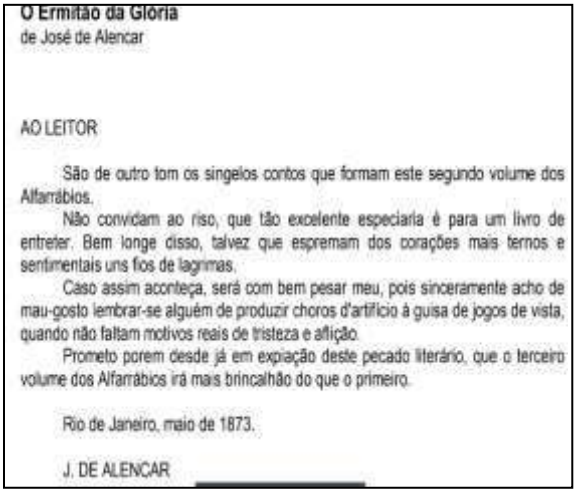
Texto-base digitalizado por:
Marciana Maria Muniz Guedes - São Paulo/SP

Este material pode ser redistribuído livremente, desde que não seja alterado, e que as informações acima sejam mantidas. Para maiores informações, escreva para <bibvirt@futuro.usp.br>.

Estamos em busca de patrocinadores e voluntários para nos ajudar a manter este projeto. Se você quer ajudar de alguma forma, mande um e-mail para <bibvirt@futuro.usp.br> e saiba

Fonte: Ao correr da pena

Figura 9: Elementos a serem removidos dos *corpora* de estudo e referência



O Ermitão da Glória
de José de Alencar

AO LEITOR

São de outro tom os singelos contos que formam este segundo volume dos Alfarrábios.

Não convidam ao riso, que tão excelente especiaria é para um livro de entreter. Bem longe disso, talvez que espremam dos corações mais ternos e sentimentais uns fios de lagrimas.

Caso assim aconteça, será com bem pesar meu, pois sinceramente acho de mau-gosto lembrar-se alguém de produzir choros d'artifício à guisa de jogos de vista, quando não faltam motivos reais de tristeza e aflição.

Prometo porem desde já em expiação deste pecado literário, que o terceiro volume dos Alfarrábios irá mais brincalhão do que o primeiro.

Rio de Janeiro, maio de 1873.

J. DE ALENCAR

Fonte: O Ermitão da Glória

Os dados apresentados nas Figuras 8 e 9 demonstram uma parcela do que foi extraído das obras do autor, dentre eles, carta ao leitor, endereço da universidade que disponibilizou o PDF do livro e número de página. O mesmo procedimento de limpeza ocorreu com todas as obras de Alencar, tanto com as que compõem o *corpus* de estudo quanto com as que compõem o *corpus* de referência.

Como o interesse desta pesquisa são os elementos que compõem o que chamaremos de corpo dos textos[33], como procedimento para limpeza foi retirado dos *corpora* tudo o que não pertencia ao corpo dos textos. Assim, os elementos retirados dos textos que compõem os *corpora* desta pesquisa foram: todos os elementos que compõem a capa; números de páginas; números do capítulos tanto em algarismos arábicos quanto em romanos; cartas ao leitor, advertências escritas pelo próprio autor; apresentações; notas de rodapé e notas de final de textos; *site* da universidade que foi responsável pela disponibilização em PDF, como www.nead.unama.br, por exemplo; palavras como FIM; índices; informações sobre coleção, nomes de organizadores e comissão de publicação; ficha catalográfica; informações sobre o autor e a obra; biografias; notas editoriais; todas as imagens; títulos como ato I, cena I, Comédia em dois atos e sequências; todas as informações adicionais, quando houve.

[33] Entende-se como corpo dos textos apenas o que é pertencente especificamente às narrativas dos romances ou às peças de teatro e às cartas. Excetua-se então todos os tipos de notas, comentários e demais elementos.

Como o recurso de introduzir notas foi explorado por Alencar, ele traz uma seção ao final do livro *Iracema* que ele próprio denominou de notas explicativa. Nela constam dos argumentos históricos, nos quais o autor traça, na sua perspectiva, um panorama para um possível surgimento do Ceará, relacionando-o com o tempo dos primeiros contatos. Porém, o autor não se estende muito neste feito, uma vez que não é seu propósito escrever uma obra histórica, mas um romance literário com fundamentos históricos, por isso ele chama de lenda para uma a formação do Ceará, conforme o trecho "Este é o argumento histórico da lenda". Continuando as notas, o autor explicita o significado de algumas expressões como

> **Quebrar a flecha** — Era entre os indígenas a maneira simbólica de estabelecer a paz entre as diversas tribos, ou mesmo entre dois guerreiros inimigos. Desde já advertimos que não se estranhe a maneira por que o estrangeiro se exprime falando com os selvagens; ao seu perfeito conhecimento dos usos e língua dos indígenas, e sobretudo a ter-se conformado com eles ao ponto de deixar os trajos europeus e pintar-se, deveu Martim Soares Moreno a influência que adquiriu entre os índios do Ceará. (Alencar, 1965, p. 88)

O autor entremeia as explicações das expressões com a descrição de algumas palavras que, na sua perspectiva necessitaria de auxílio para compreensão do leitor da época, como na nota para a palavra "**U**iraçaba — Aljava, de *uira* — seta, e a desinência *çaba* — coisa própria". No corpo do texto do romance, o vocábulo é utilizado no período "O sentimento que ele pôs nos olhos e no rosto, não o sei eu. Porém a virgem lançou de si o arco e a uiraçaba, e correu para o guerreiro, sentida da mágoa que causara (ALENCAR, 1965, p. 05).

Observamos que, neste caso, o autor não apresenta o significado do vocábulo *uiraçaba* na nota, mas o seu processo de composição e cria um étimo. Ao analisar o contexto de emprego do vocábulo *uiraçaba* e o seu processo de formação, é possível perceber que a nota do autor não esclarece suficientemente o significado do vocábulo empregado no contexto transcrito. É necessário que o leitor lance mão do recurso de inferência para compreender que se trata de um instrumento, provavelmente um tipo de flecha, utilizadas pelos indígenas. Dessa forma, as notas do autor auxiliam na compreensão dos termos, porém não são suficientes.

Há, porém, algumas palavras nas notas, cujas definições parecem compatíveis e suficientes com o contexto de emprego como

> **Guaiúba –** de goiá – vale, y – água, jur – vir, be – por onde: por onde vem as águas do vale. Rio que nasce na serra da Aratanha e corta a povoação do mesmo nome a seis léguas da capital (ALENCAR, 1965, p. 158)

Ao se comparar com o contexto de emprego em

> Uma manhã Poti guiou Martim à caça. Caminharam para uma serra, que se levanta ao lado da outra do Maranguab, sua irmã. O alto cabeço se curva à semelhança do bico adunco da arara; pelo que os guerreiros a chamaram Aratanha. Eles subiram pela encosta da Guaiúba por onde as águas descem para o vale, e foram até o córrego habitado pelas pacas. (ALENCAR, 1965, p. 111)

é possível dizer que a definição apresentada pelo autor na nota é suficiente para a compreensão do significado do vocábulo. Nos casos como esse em que a definição do autor for suficiente para elucidação do significado do item lexical, ela será utilizada para a definição do item lexical no Vocabulário.

Após a análise contrastiva dos vocábulos apresentados nas notas de Alencar com o contexto de emprego, ficou evidente que uma análise descritiva baseada nos contextos de emprego poderia ser mais producente para os leitores, o que viabilizaria o Vocabulário que sugerimos organizar ao final desta pesquisa. Isso pela razão de que o autor não traz o significado de todos os vocábulos indígenas empregadas nas três obras, e quando o faz, em muitos casos, o significado não é suficiente para a compreensão do item lexical no contexto.

Como haveria a possibilidade de consulta às notas do autor para elaboração da descrição do item lexical, os livros em PDF, na íntegra, foram mantidos e armazenados, como já demonstrado.

Finalizado o processo de preparação do *corpus* de estudo, a próxima etapa consiste na extração das listas de palavras para o levantamento dos dados estatísticos, conforme demonstramos na seção seguinte.

4.2.4 Extração das listas de palavras para o levantamento dos dados estatísticos

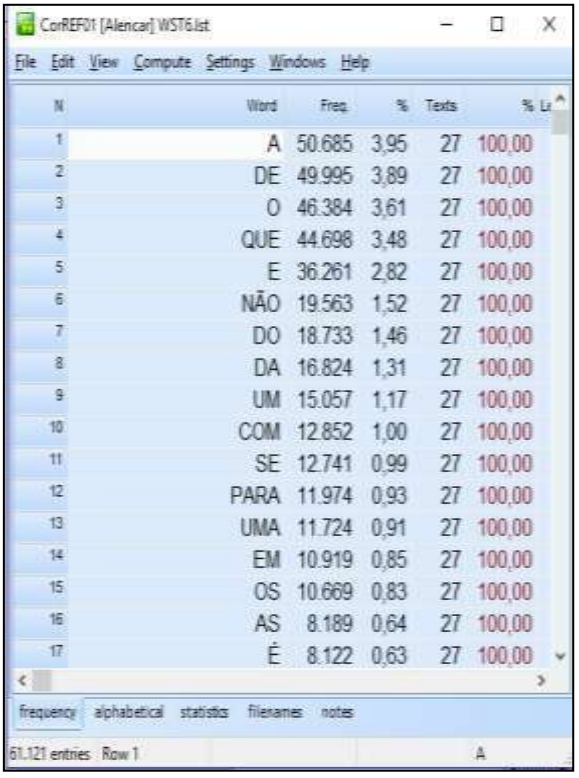
Para que um *corpus* possa ser processado pelo Programa *WST*, o primeiro passo é gerar uma lista de palavras, por meio da ferramenta *WordList*. A partir desse processo, levantamos os dados estatísticos dos *corpora* com as listas de todas as palavras do *corpus* de

estudos e, em seguida, do *corpus* de referência. A Figura 10 apresenta as 17 primeiras palavras do *corpus* de estudo e a Figura 11, as 17 primeiras palavras do *corpus* de referência, com as respectivas frequências, porcentagem de ocorrências em relação à totalidade do *corpus* e número de texto em que ocorrem. É possível notar que todas as palavras das listas aparecem em todos os textos: no *corpus* de estudos, os três textos: *Iracema*, *Ubirajara* e *O Guarani* e, no *corpus* de referência, as 27 obras que o compõem.

Figura 10: Lista das 17 primeiras palavras do *corpus* de estudo

Fonte: A autora, por meio da Ferramenta *WordList* do *WST* 6.0.

Figura 11: Lista das 17 primeiras palavras do *corpus* de referência de Alencar



Fonte: A autora, por meio da ferramenta *WordList* do *WST* 6.0.

É possível visualizar que as palavras que estão no topo das mais frequentes são palavras gramaticais como artigos, pronomes e contrações.

Vale ressaltar que já possuímos, neste momento, as listas de palavras dos outros três *corpora* que também serão utilizados para comparação com o *corpus* de estudo de Alencar. Esses três *corpora* são: *corpus* CorpRef-Lácio-Web; CorpRef-AcadTeses; CorpRef-No, além

do já mencionado *corpus* de referência formado pelas obras não indianistas do próprio autor José de Alencar, o CorpRef-Alencar.

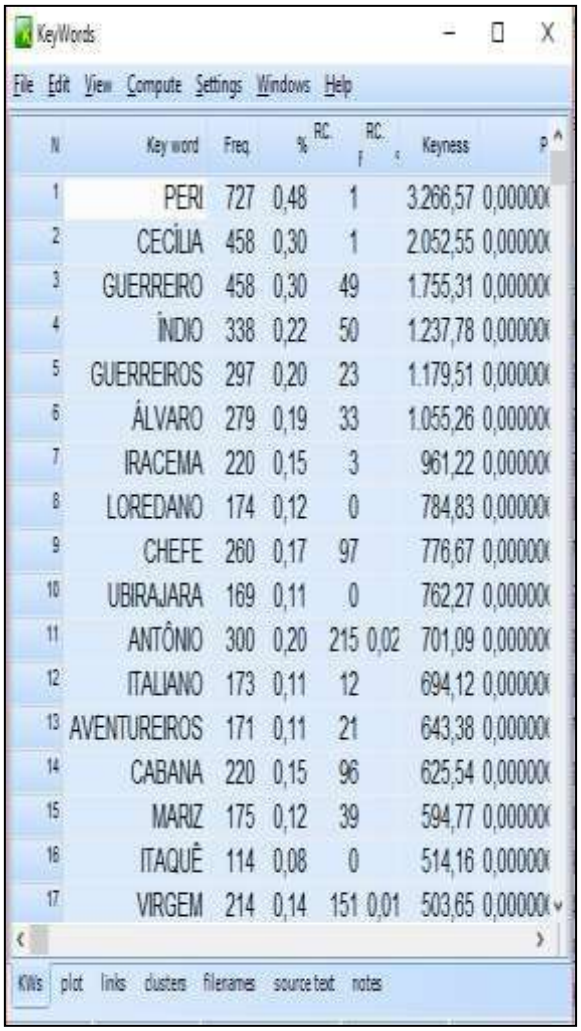
Ressaltamos que as listas de palavras dos outros três *corpora* de referência que serão utilizados para a comparação com o *corpus* de estudo de Alencar. Esses três *corpora* são: *corpus* CorpRef-Lácio-Web, CorpRef-AcadTeses e CorpRef-Nov, além do *corpus* de referência formado pelas obras não indianistas do próprio autor José de Alencar, o CorpRef-Alencar.

Após a extração das listas de palavras para cada um dos *corpora* e salvas em formato *lst* (*WordList*), procedemos ao levantamento das palavras-chave, cujos procedimentos detalhamos na seção seguinte.

4.2.5 Extração das listas de palavras-chave

Após a realização dos passos para obtermos as listas de palavras de cada um dos *corpora*, tanto de estudo quanto de referência, seguimos os procedimentos para obtenção das palavras-chave por meio da ferramenta *KeyWords.* A comparação foi realizada com cada um dos *corpora* de referência. A Figura 12 apresenta as palavras-chave resultantes do contraste entre o *corpus* de estudo e o *corpus* de referência CorpRef-Alencar e a Figura 13 apresenta o contraste com o *corpus* de referência CorpRef-Lácio-Web.

Figura 12: Palavras-chave com CorpRef-Alencar



KeyWords

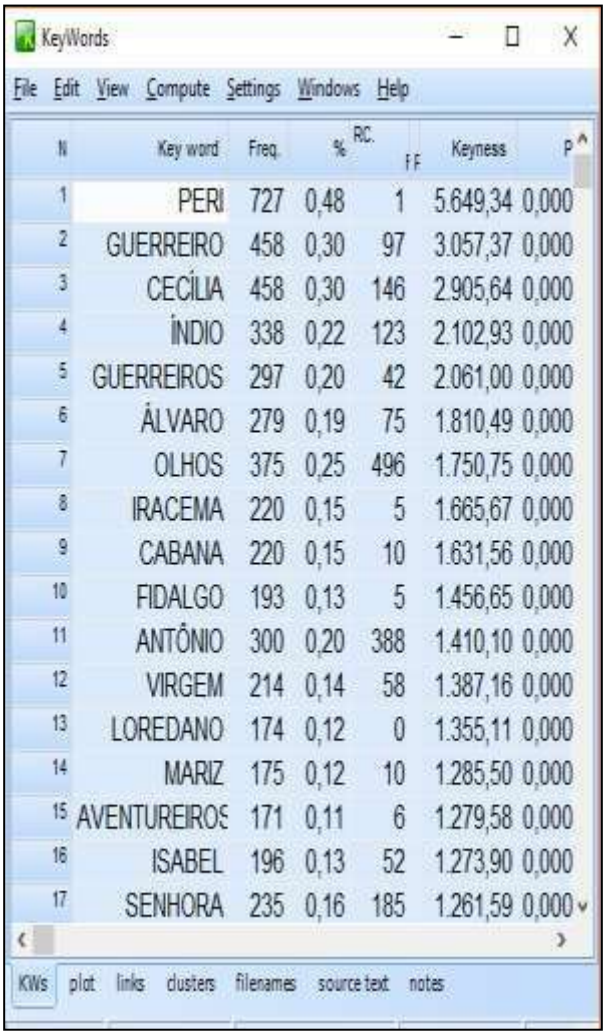
File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	RC.	RC. F	Keyness	P
1	PERI	727	0,48	1		3.266,57	0,00000
2	CECÍLIA	458	0,30	1		2.052,55	0,00000
3	GUERREIRO	458	0,30	49		1.755,31	0,00000
4	ÍNDIO	338	0,22	50		1.237,78	0,00000
5	GUERREIROS	297	0,20	23		1.179,51	0,00000
6	ÁLVARO	279	0,19	33		1.055,26	0,00000
7	IRACEMA	220	0,15	3		961,22	0,00000
8	LOREDANO	174	0,12	0		784,83	0,00000
9	CHEFE	260	0,17	97		776,67	0,00000
10	UBIRAJARA	169	0,11	0		762,27	0,00000
11	ANTÔNIO	300	0,20	215	0,02	701,09	0,00000
12	ITALIANO	173	0,11	12		694,12	0,00000
13	AVENTUREIROS	171	0,11	21		643,38	0,00000
14	CABANA	220	0,15	96		625,54	0,00000
15	MARIZ	175	0,12	39		594,77	0,00000
16	ITAQUÊ	114	0,08	0		514,16	0,00000
17	VIRGEM	214	0,14	151	0,01	503,65	0,00000

KWs plot links clusters filenames source text notes

Fonte: A autora, por meio do *WST*

Figura 13: Palavras-chave com o CorpRef-Lácio-Web



KeyWords

File Edit View Compute Settings Windows Help

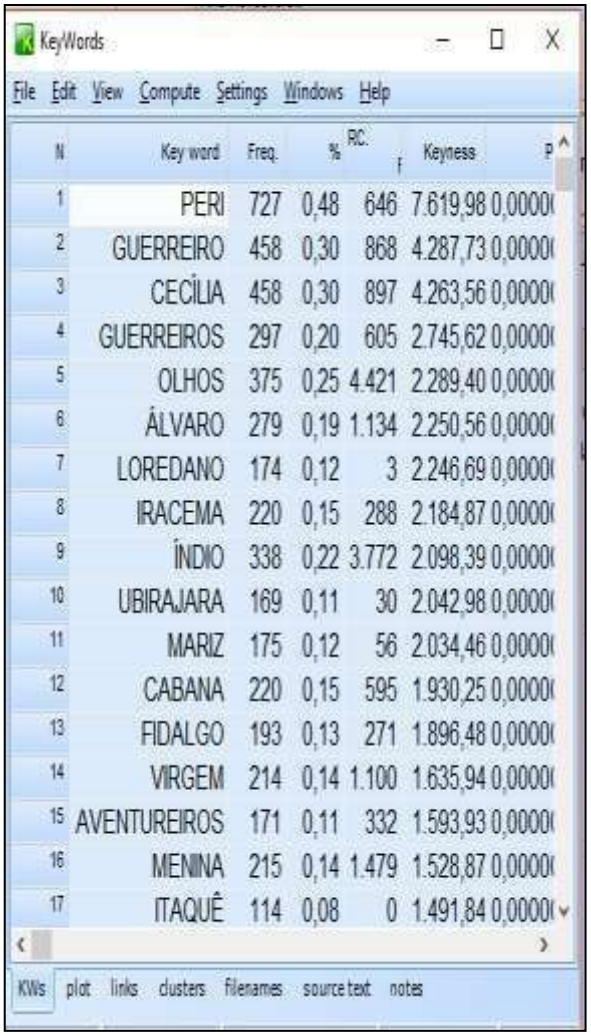
N	Key word	Freq.	%	RC.	RC. F F	Keyness	P
1	PERI	727	0,48	1		5.649,34	0,000
2	GUERREIRO	458	0,30	97		3.057,37	0,000
3	CECÍLIA	458	0,30	146		2.905,64	0,000
4	ÍNDIO	338	0,22	123		2.102,93	0,000
5	GUERREIROS	297	0,20	42		2.061,00	0,000
6	ÁLVARO	279	0,19	75		1.810,49	0,000
7	OLHOS	375	0,25	496		1.750,75	0,000
8	IRACEMA	220	0,15	5		1.665,67	0,000
9	CABANA	220	0,15	10		1.631,56	0,000
10	FIDALGO	193	0,13	5		1.456,65	0,000
11	ANTÔNIO	300	0,20	388		1.410,10	0,000
12	VIRGEM	214	0,14	58		1.387,16	0,000
13	LOREDANO	174	0,12	0		1.355,11	0,000
14	MARIZ	175	0,12	10		1.285,50	0,000
15	AVENTUREIROS	171	0,11	6		1.279,58	0,000
16	ISABEL	196	0,13	52		1.273,90	0,000
17	SENHORA	235	0,16	185		1.261,59	0,000

KWs plot links clusters filenames source text notes

Fonte: A autora, por meio do *WST*

A Figura 14 apresenta o contraste entre o *corpus* de estudo de Alencar com o *corpus* CorpRef-AcadTeses e a Figura 15 com o *corpus* CorpRef-Nov.

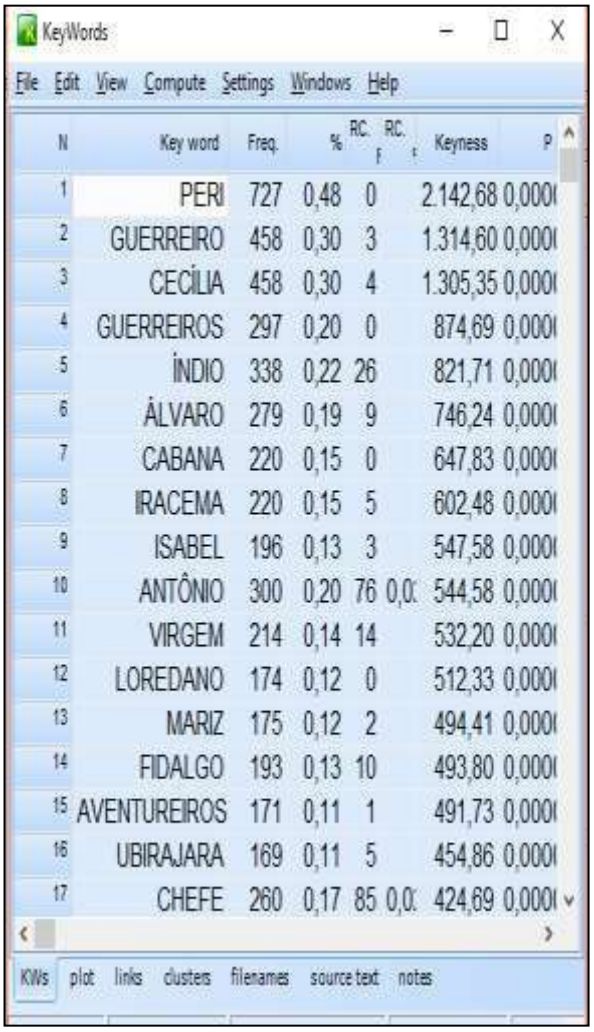
Figura 14: Palavras-chave com o CorpRef-AcadTeses



N	Key word	Freq.	%	RC. f	Keyness	P
1	PERI	727	0,48	646	7.619,98	0,0000
2	GUERREIRO	458	0,30	868	4.287,73	0,0000
3	CECÍLIA	458	0,30	897	4.263,56	0,0000
4	GUERREIROS	297	0,20	605	2.745,62	0,0000
5	OLHOS	375	0,25	4.421	2.289,40	0,0000
6	ÁLVARO	279	0,19	1.134	2.250,56	0,0000
7	LOREDANO	174	0,12	3	2.246,69	0,0000
8	IRACEMA	220	0,15	288	2.184,87	0,0000
9	ÍNDIO	338	0,22	3.772	2.098,39	0,0000
10	UBIRAJARA	169	0,11	30	2.042,98	0,0000
11	MARIZ	175	0,12	56	2.034,46	0,0000
12	CABANA	220	0,15	595	1.930,25	0,0000
13	FIDALGO	193	0,13	271	1.896,48	0,0000
14	VIRGEM	214	0,14	1.100	1.635,94	0,0000
15	AVENTUREIROS	171	0,11	332	1.593,93	0,0000
16	MENINA	215	0,14	1.479	1.528,87	0,0000
17	ITAQUÊ	114	0,08	0	1.491,84	0,0000

Fonte: A autora, por meio do *WST*

Figura 15: Palavras-chave com o CorpRef-Nov



N	Key word	Freq.	%	RC. f	RC. r	Keyness	P
1	PERI	727	0,48	0		2.142,68	0,000
2	GUERREIRO	458	0,30	3		1.314,60	0,000
3	CECÍLIA	458	0,30	4		1.305,35	0,000
4	GUERREIROS	297	0,20	0		874,69	0,000
5	ÍNDIO	338	0,22	26		821,71	0,000
6	ÁLVARO	279	0,19	9		746,24	0,000
7	CABANA	220	0,15	0		647,83	0,000
8	IRACEMA	220	0,15	5		602,48	0,000
9	ISABEL	196	0,13	3		547,58	0,000
10	ANTÔNIO	300	0,20	76	0,0	544,58	0,000
11	VIRGEM	214	0,14	14		532,20	0,000
12	LOREDANO	174	0,12	0		512,33	0,000
13	MARIZ	175	0,12	2		494,41	0,000
14	FIDALGO	193	0,13	10		493,80	0,000
15	AVENTUREIROS	171	0,11	1		491,73	0,000
16	UBIRAJARA	169	0,11	5		454,86	0,000
17	CHEFE	260	0,17	85	0,0	424,69	0,000

Fonte: A autora, por meio do *WST*

Uma primeira observação desses resultados evidencia algumas características marcantes, por exemplo, a palavra Peri é a mais chave no contraste com todos os *corpora* de referência. Nota-se também que não houve alterações significativas nos contrastes em relação às demais primeiras 17 palavras-chave das listas, organizadas por chavicidade.

Finalizamos esta seção que teve por objetivo apresentar os procedimentos para contrastar o *corpus* de estudos com os *corpora* de referência. Uma análise mais criteriosa acerca dos contrastes ocorrerá no capítulo de análise.

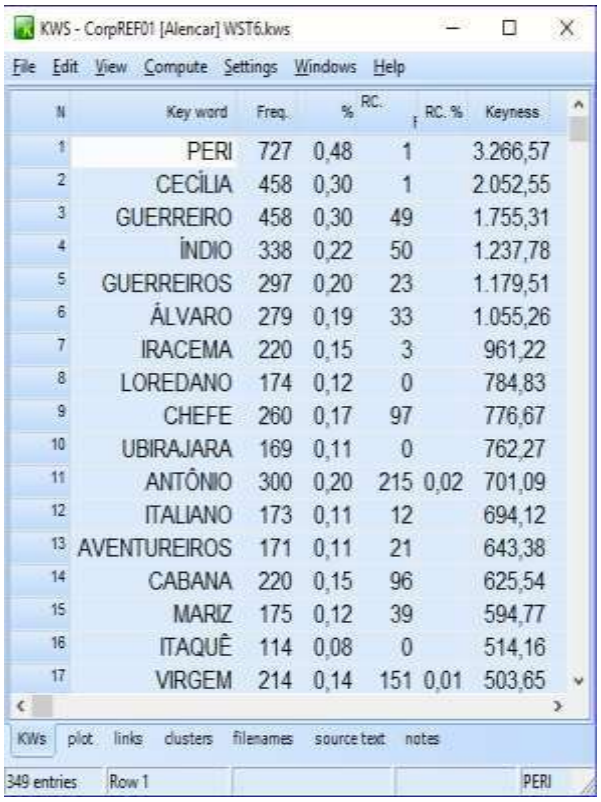
4.2.6 Análise contrastiva das diferentes listas de palavras, a partir dos diferentes **corpora** *de referência*

O levantamento das palavras-chave foi um procedimento relevante para a observação do léxico das obras indianistas de Alencar em contraste com as demais obras do mesmo autor. A ferramenta *KeyWords* possibilita a comparação entre as listas de palavras do *corpus* de estudos com as listas de palavras do *corpus* de referência, extraindo as palavras-chave.

Para além da extração das palavras-chave, nosso propósito é contrastar as quatro listas de palavras-chave entre si, a fim de verificar até que ponto o contraste de Alencar com ele mesmo seria produtivo para os fins de identificação do léxico indianista. Além disso, conferir as diferenças quanto à utilização de *corpora* de referência de diferentes extensão e representatividade. Após esse contraste, os resultados tenderiam a ser mais confiáveis, uma vez que partiram da comparação entre os resultados de quatro listas.

Nesse sentido, conforme mencionado, foram extraídas quatro listas de palavra-chave com o intuito de comparação, as quais são: i) *corpus* de estudo com o *corpus* de referência do próprio José de Alencar, CorpRef-Alencar; ii) *corpus* de estudo com o CorpRef-Lácio-Web; iii) *corpus* de estudo com o CorpRef-AcadTeses e iv) *corpus* de estudo com CorpRef-Nov. As Figuras 16 e 17 apresentam uma visão parcial das listas de palavras-chave do *corpus* de estudo em contraste com os *corpora* de referência CorpRef-Alencar e *corpus* de estudo com o CorpRef-Lácio-Web, respectivamente. As listas estão organizadas em função da chavicidade, após a leitura com a ferramenta *KeyWords*. As demais listas de palavras estão armazenadas no computador pessoal e na plataforma *Dropbox* para consulta durante as análises.

Figura 16: Visão parcial da lista de palavras-chave: *corpus* de estudo com CorpRef-Alencar



N	Key word	Freq.	%	RC.	RC. %	Keyness
1	PERI	727	0,48	1		3.266,57
2	CECÍLIA	458	0,30	1		2.052,55
3	GUERREIRO	458	0,30	49		1.755,31
4	ÍNDIO	338	0,22	50		1.237,78
5	GUERREIROS	297	0,20	23		1.179,51
6	ÁLVARO	279	0,19	33		1.055,26
7	IRACEMA	220	0,15	3		961,22
8	LOREDANO	174	0,12	0		784,83
9	CHEFE	260	0,17	97		776,67
10	UBIRAJARA	169	0,11	0		762,27
11	ANTÔNIO	300	0,20	215	0,02	701,09
12	ITALIANO	173	0,11	12		694,12
13	AVENTUREIROS	171	0,11	21		643,38
14	CABANA	220	0,15	96		625,54
15	MARIZ	175	0,12	39		594,77
16	ITAQUÊ	114	0,08	0		514,16
17	VIRGEM	214	0,14	151	0,01	503,65

Fonte: A autora, por meio da ferramenta *KeyWord* do *WST*

Figura 17: Visão parcial da lista de palavras-chave: *corpus* de estudo com CorpRef-Lácio-Web

N	Key word	Freq.	%	RC.	RC.	Keyness
1	PERI	727	0,48	1		5.649,34
2	GUERREIRO	458	0,30	97		3.057,37
3	CECÍLIA	458	0,30	146		2.905,64
4	ÍNDIO	338	0,22	123		2.102,93
5	GUERREIROS	297	0,20	42		2.061,00
6	TU	297	0,20	52		2.021,57
7	ÁLVARO	279	0,19	75		1.810,49
8	SEU	1.150	0,76	11.02	0,15	1.800,03
9	OLHOS	375	0,25	496		1.750,75
10	IRACEMA	220	0,15	5		1.665,67
11	CABANA	220	0,15	10		1.631,56
12	TE	270	0,18	161		1.539,91
13	TINHA	493	0,33	1.822	0,03	1.517,63
14	FIDALGO	193	0,13	5		1.456,65
15	QUE	5.537	3,68	152.0	2,10	1.453,64
16	ANTÔNIO	300	0,20	388		1.410,10

Fonte: A autora, por meio da ferramenta *KeyWord* do *WST*

As listas de palavras, por si sós, não possibilitam muitas considerações a respeito do léxico indianista, porém o contraste da lista de palavras do *corpus* de estudo com os *corpora* de referência, por meio das palavras-chave, já nos permite uma visão do perfil dos elementos linguísticos chave do *corpus* de estudo. No caso das duas listas de palavras-chave geradas, por meio da ferramenta *KeyWords*, é possível notar que há um número significativo de palavras-chave que se repetem nas duas listas, como **Peri**, **Cecília**, **guerreiro**, **Iracema**, **cabana**, dentre outros.

Para obtenção das listas de palavras-chave, utilizamos o valor de significância estatística *p*= 0,0000001, que é *defaut* do programa, cuja probabilidade de resultados a serem obtidos por acaso é de 1 em 1 milhão. Segundo Berber Sardinha (2009), quanto menor for o número de p, maior é a significância; e frequência foi alterada para mínimo 2. Ao contrastar as listas de palavras-chave, obtivemos os seguintes resultados: com o *corpus* de referência CorpRef-Alencar, o resultado foram 349 palavras-chave, das quais apresentamos as 17 primeiras, conforme Figura 16; com o *corpus* de referência CorpRef-Lácio-Web obtivemos 1.724 palavras-chave, conforme Figura 17, com a qual apresentamos as 16 primeiras; com o *corpus* de referência CorpusRef-Alencar, obtivemos 349 palavras-chave; com o *corpus* de referência CorpRef-AcadTeses, obtivemos 2.080 palavras-chave. A palavra Peri foi a mais a mais chave em todos os *corpora* de referência[34].

De posse das listas de palavras-chave salvas em formato *KWS*, procedemos à organização dos itens lexicais, com o propósito de identificar os candidatos a vocábulos indianistas de José de Alencar. A seguir apresentamos os procedimentos adotados para essa organização.

4.2.7 Identificação do léxico indianista

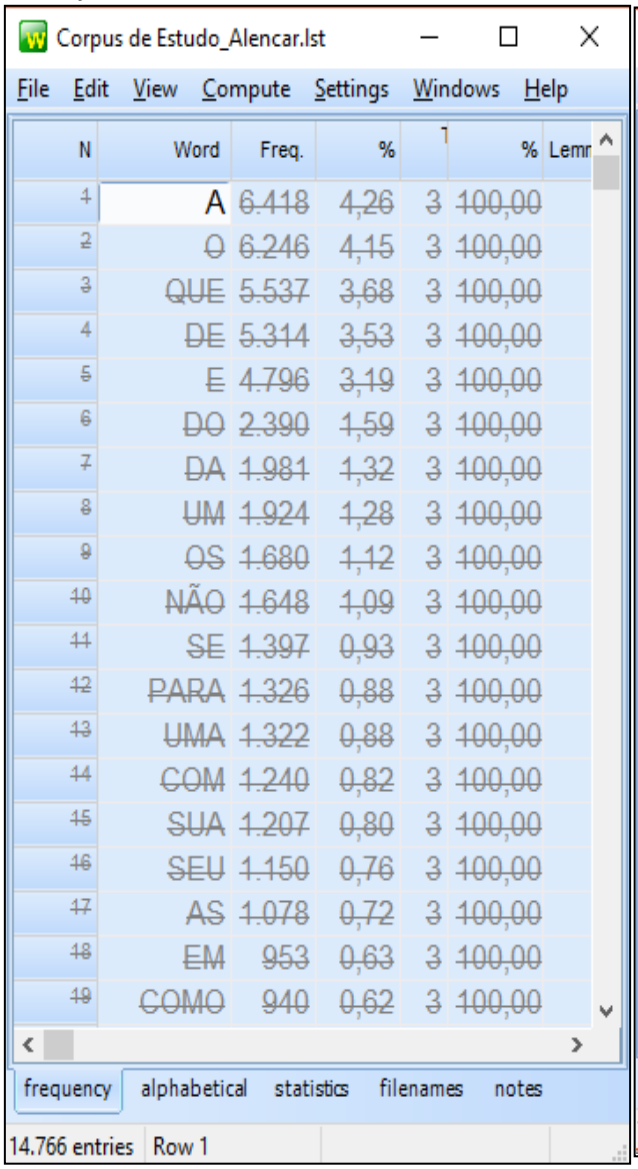
Para identificar o léxico indianista de Alencar nas obras *Iracema*, *Ubirajara* e *O Guarani*, alguns procedimentos foram adotados. Após a extração e limpeza das palavras-chave, em decorrência da percepção de que alguns vocábulos indianistas pudessem não ter sido capturados pela ferramenta, por segurança, também foi realizada uma conferência nas listas de palavras do *corpus* de estudo. Isso foi também em virtude de cuidar para que dados importantes não passassem desapercebidos, assim, paralelamente, ao estudo contrastivo e limpeza das listas de palavras-chave foi realizado um estudo junto à lista de palavras do *corpus* de estudo.

Em seguida, iniciamos o trabalho manual de exclusão dos itens gramaticais e das formas verbais e, para facilitar o processo de limpeza, reordenamos em ordem de frequência para procedermos ao primeiro passo, que foi a exclusão das palavras gramaticais e das formas verbais. A exclusão das formas verbais se deveu pelo fato de não ser objetivo da pesquisa, já

[34] Será realizada uma análise minuciosa sobre esses resultados no Capítulo 5 "Análise contrastiva do *corpus* de estudo de Alencar com os *corpora* de referência".

que demandaria outro estudo em razão da quantidade de formas verbais oriundas da conjugação de um mesmo verbo. A *WordList* era composta de 14.766 itens, conforme demonstrado na Figura 18. O procedimento de exclusão foi um trabalho exaustivo de aproximadamente dezesseis horas. Após a exclusão das palavras gramaticais e das formas verbais restaram 758 itens que chamamos de candidatos a vocábulos indianistas, como demonstramos na Figura 19.

Figura 18: Vista parcial do procedimento de extração do léxico indianista



Fonte: A autora, por meio do *WST*

Figura 19: Candidatos a vocábulos indianistas

Fonte: A autora, por meio do *WST*

Esta lista de palavra, após excluídos os itens gramaticais e as formas verbais, foi armazenada e procedemos a uma etapa seguinte para extração do léxico indianista. Como a etimologia das palavras não era conhecida, optamos por uma primeira consulta de todas as palavras da lista, no dicionário *Houaiss* versão eletrônica 2009. A consulta se justifica, na medida em que adotamos como critério o fato de que se o *Houaiss* apontasse a origem da palavra, e ela não fosse de origem indígena, seria excluída da lista, conforme apresentado na Figura 20. Mantivemos, neste momento, todas as palavras adotando os critérios: palavras, cuja origem não fosse explicitada pelo dicionário *Houaiss* inclusive os nomes próprios; palavras não constantes da lista dos verbetes do dicionário; e as palavras declaradas de origem indígena pelo dicionário.

Figura 20: Vista parcial das palavras não indígenas excluídas da lista de palavras

Fonte: A autora, por meio da ferramenta *WordList*

Como critério de explicitação, apresentamos como estas palavras marcadas na Figura 20 são definidas pelo dicionário *Houaiss*, 2009:

Guerreiro – guerra – germ. ocidental *werra* ‘discórdia, revolta, peleja’;

Chefe – fr. *chef* ‘aquele que está à frente de qualquer coisa’ Italiano – top. *Itália* + -ano;

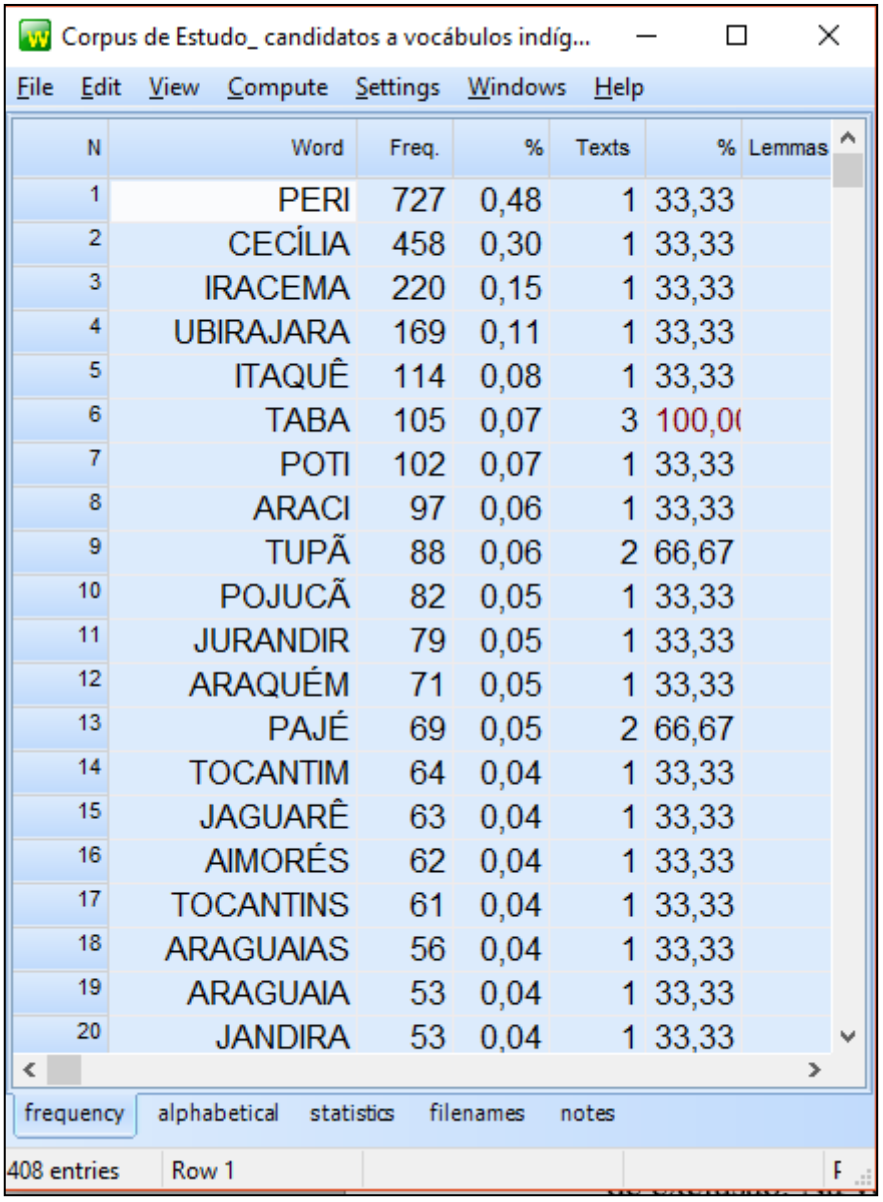
Italiano – top. Itália + -ano

Armas – lat. arma, ae ‘arma’, no neutro pl amorum ‘armas em geral’

Podemos notar que, na perspectiva do dicionário consultado, as palavras apresentadas na lista não são de origem indígena, fato que nos levou a excluí-las. Esse mesmo procedimento de consulta foi adotado com todas as demais palavras da lista.

Com base nessa listagem de candidatos, foi feita uma análise manual, a fim de determinar os casos verdadeiros de vocábulo indianista. Embora possamos inferir que palavras como Álvaro e Martim também não pertençam ao indianismo, optamos por mantê-las na lista até que fizéssemos novas pesquisas em outros dicionários para garantir a certeza da dedução. Assim, neste momento, somente extraímos as palavras que figuraram no dicionário *Houaiss* com a etimologia claramente definida, como nos exemplos apresentados, como é o caso de ‘chefe’, por exemplo, que é de origem francesa. Desta etapa da pesquisa restaram 408 itens, como demonstra a Figura 21.

Figura 21: Lista de candidatos a vocábulos após consulta ao *Houaiss*



N	Word	Freq.	%	Texts	%	Lemmas
1	PERI	727	0,48	1	33,33	
2	CECÍLIA	458	0,30	1	33,33	
3	IRACEMA	220	0,15	1	33,33	
4	UBIRAJARA	169	0,11	1	33,33	
5	ITAQUÊ	114	0,08	1	33,33	
6	TABA	105	0,07	3	100,0(	
7	POTI	102	0,07	1	33,33	
8	ARACI	97	0,06	1	33,33	
9	TUPÃ	88	0,06	2	66,67	
10	POJUCÃ	82	0,05	1	33,33	
11	JURANDIR	79	0,05	1	33,33	
12	ARAQUÉM	71	0,05	1	33,33	
13	PAJÉ	69	0,05	2	66,67	
14	TOCANTIM	64	0,04	1	33,33	
15	JAGUARÊ	63	0,04	1	33,33	
16	AIMORÉS	62	0,04	1	33,33	
17	TOCANTINS	61	0,04	1	33,33	
18	ARAGUAIAS	56	0,04	1	33,33	
19	ARAGUAIA	53	0,04	1	33,33	
20	JANDIRA	53	0,04	1	33,33	

Fonte: A autora, por meio do *WST*

Após os procedimentos descritos, iniciamos uma busca pelas palavras nos dicionários de consulta. Pesquisamos, então, as 408 palavras nos dicionários de língua indígena:

i) *Dicionário de língua Tupy*, de A. Gonçalves Dias (1858) - GD;

ii) *O tupi da geografia nacional***,** de Theodoro Sampaio (1901) – TS;

iii) *Dicionário Histórico das palavras Portuguesas de origem Tupi***,** de Antônio Geraldo da Cunha (1998) – AGC.

iv) *O Dicionário Tupi-português***,** de Luiz Caldas Tibiriçá (1984) - LCT

Utilizamos também, os dicionários de língua geral do português que datam de um período anterior às publicações de José de Alencar.[35]

v) *Dicionário da Lingua Brasileira*, de Luiz Maria da Silva Pinto (1832) – LMSP
vi) *Diccionário da Lingua Portugueza*, de Antonio de Moraes Silva (1789) – AMS e
vii) *Vocabulario Portuguez e Latino*, de Raphael Bluteau (1712) - RB

Com a finalidade de sistematização e facilitação da leitura, os dicionários poderão ser referenciados por meio das iniciais, conforme proposto no Quadro 6.

Quadro 6: Sistematização dos dicionários

Sigla	***Dicionário, autor e ano de publicação***
AGC	*Dicionário Histórico das palavras Portuguesas de origem Tupi,* de Antônio Geraldo da Cunha (1998)
LCT	*O Dicionário Tupi-português,* de Luiz Caldas Tibiriçá (1984)
TS	*O tupi da geografia nacional,* de Theodoro Sampaio (1901)
GD	*Dicionário de língua Tupy,* de A. Gonçalves Dias (1858)
LMSP	*Dicionário da Lingua Brasileira*, de Luiz Maria da Silva Pinto (1832)
AMS	*Diccionário da Lingua Portugueza*, de Antonio de Moraes Silva (1789)
RB	*Vocabulario Portuguez e Latino,* de Raphael Bluteau (1712)

Fonte: A autora

Tomemos como exemplo dos procedimentos o primeiro vocábulo da lista, **Peri**. Pesquisamos, inicialmente, e como mencionado, no dicionário Houaiss (2009), que apresenta a definição "m.q. piri (angicos, 'terreno', 'vegetação'", mas não traz a formação ou

[35] As versões desses dicionários foram disponibilizadas, em PDF, pela Biblioteca Brasiliana da USP,

etimologia da palavra, por isso a mantivemos na lista de candidatos a vocábulos. Em seguida, recorremos aos dicionários de exclusão e obtivemos a seguintes definições:

GD – s. povoação do Maranhão. De *pery* ou *piry*, junco. Existe Peri de Cima e Peri de Baixo, separados por pequenos espaços de terra.

TS – Pery corr. *pary* ou *piri*, o junco. (p. 146)

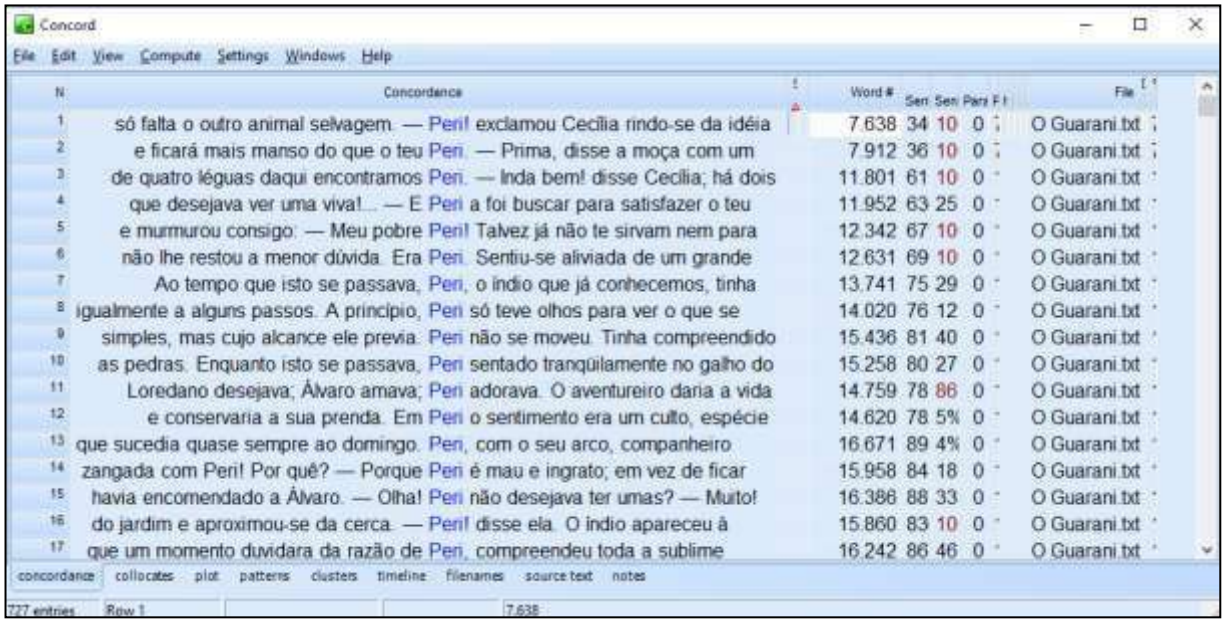
Nos demais: **AGC, LCT, MAS**, **RB, AMS** e **LMSP** – não consta o vocábulo.

É possível notar que apenas dois dos dicionários consultados trazem a palavra **Peri**, porém nenhum faz menção a **Peri** referindo-se a antropônimo, como é utilizado pelo autor José de Alencar no livro O *Guarani,* para nomear seu personagem principal e representante indígena. Como os dicionários GD e TS fazem referência a **junco** fizemos uma busca no Houaiss eletrônico (2009) da palavra **junco**, a qual é descrita como "design. comum às ervas do gên. *juncus*, da fam. das juncáceas, que reúne certa de 300 spp". Com base nos dicionários consultados, **peri** é o nome de uma planta, porém Alencar a utiliza como antropônimo. Alencar não define **Peri**, porém compara as características do índio com o junco selvagem, conforme o trecho da obra *O Guarani* "Uma simples túnica de algodão, a que os indígenas chamavam aimará, apertada à cintura por uma faixa de penas escarlates, caía-lhe dos ombros até ao meio da perna, e desenhava o talhe delgado e esbelto como um junco selvagem". É possível notar que Alencar estabelece uma relação entre as características físicas do seu personagem com as da árvore da qual toma emprestado o nome.

Outra fonte para a pesquisa sobre os vocábulos de Alencar serão as linhas de concordância, as quais foram utilizadas para elaboração da descrição dos vocábulos empregados por Alencar nas obras indianistas. No livro *O Guarani,* **Peri** nomeia o índio, personagem principal da narrativa. Apesar das críticas de que Alencar não retratara a realidade, ele criou uma ficção com base na imaginação e nos estudos sobre os índios que fizera, como afirma Schwamborn (1998, p. 374) "Alencar hauriu suas ideias a respeito dos índios nas fontes históricas e literárias que estimularam sua fantasia". O nome escolhido para seu índio expressa o desejo do autor em caracterizá-lo como selvagem, como lhe é característico, e "talhe delgado e altivo", ou seja, a imagem que atende ao seu desejo

“selvagem, alto e postura de herói”. A Figura 22 apresenta uma visão parcial das linhas de concordância do vocábulo **Peri**.

Figura 22: Linhas de concordância da palavra **Peri**



Fonte: A autora, por meio da ferramenta *Concord* do *WST*

Embora os nomes próprios não sejam, costumeiramente, definidos, no caso das obras de Alencar, eles são passíveis de descrição, tendo em vista que o autor os cria, em grande parte, com o propósito de atender aos seus desejos em relação ao personagem, ou seja, os antropônimos assumem um papel mais importante do que a identificação dos personagens para ser a própria descrição das características desses personagens.

Após este procedimento para abstrair o léxico indianista, obtivemos uma lista de 367 itens indianistas, os quais comporão o vocabulário indianista de José de Alencar. A Figura 23 mostra as 20 primeiras palavras indianistas por ordem de frequência.

Figura 23: Lista do léxico indianista de José de Alencar

N	Word	Freq.	%		%	Len
1	PERI	727	0,48	1	33,33	
2	IRACEMA	220	0,15	1	33,33	
3	UBIRAJARA	169	0,11	1	33,33	
4	ITAQUÊ	114	0,08	1	33,33	
5	TABA	105	0,07	3	100,0	
6	POTI	102	0,07	1	33,33	
7	ARACI	97	0,06	1	33,33	
8	TUPÃ	88	0,06	2	66,67	
9	POJUCÃ	82	0,05	1	33,33	
10	JURANDIR	79	0,05	1	33,33	
11	ARAQUÉM	71	0,05	1	33,33	
12	PAJÉ	69	0,05	2	66,67	
13	TOCANTIM	64	0,04	1	33,33	
14	JAGUARÊ	63	0,04	1	33,33	
15	AIMORÉS	62	0,04	1	33,33	
16	TOCANTINS	61	0,04	1	33,33	
17	ARAGUAIAS	56	0,04	1	33,33	
18	ARAGUAIA	53	0,04	1	33,33	
19	JANDIRA	53	0,04	1	33,33	
20	TABAJARAS	51	0,03	1	33,33	

Fonte: A autora, por meio do *WST*

Após a descrição dos passos que utilizamos para chegar ao léxico indianista do autor, agrupamos em campos semânticos, cujos critérios serão descritos na próxima seção.

4.2.8 Agrupamento do léxico indianista por campos semânticos

Após as pesquisas e análises seguindo os procedimentos descritos na seção "Identificação do léxico indianista", tornou possível realizar as primeiras análises dos vocábulos e estratificá-los em campos semânticos. Para a separação por grupos semânticos, adotamos as definições propostas pelos dicionários de exclusão e as linhas de concordância geradas a partir da ferramenta *Concord* do *WST* 6.0. Assim definimos dez campos semânticos: antropônimos, topônimos, Etnônimos, flora, fauna, utensílios e vestimentas, alimentos, espiritualidade e campo semântico denominamos diversos, no qual inserimos os vocábulos que não se encaixavam nos anteriores e que não são em número significativo para um novo grupo semântico. O Quadro 7 apresenta os vocábulos distribuídos nos campos semânticos. A totalidade do léxico distribuído em campos semânticos será apresentada no Capítulo 07 "Léxico indianista de José de Alencar: proposta para um Vocabulário *online"*, trazemos, portanto, uma amostra da distribuição.

Quadro 7: Léxico indianista distribuído em campos semânticos

Antropônimos	Etnônimos	Topônimos	Flora	Fauna
- Abaeté - Aimoré - Andira - Araci - Araquém - Ararê - Araribóia - Aresqui - Batuirité - Boitatá	- Aimoré - Araguaia - Caraíba - Caramuru - Emboabas - Goitacá - Guaraciaba - Guarani - Marabá - Moacara	- Acaraú - Acarú - Apodi - Aratanha - Aratuba - Arroio - Campina - Camucim - Capoeira - Ceará	- Abacaxi - Acajás - Açaí - Acaris - Aguapés - Aipim - Airi - Ananás - Andiroba - Angico	- Acauã - Anajê - Andira - Ará - Arara - Arirama - Ariranha - Araruna - Ati - Atiati
Utensílios/ vestimentas	Alimentos	Espiritualidade	Habitação	Diversos

- Araçóia - Boré - Caiçara - Camuci - Camucim - Carioba - Clavina - Igaçaba - Igara - Inúbia	- Abati - Biaribi - Carimã - Cauim - Jurema - Mingau - Moquém - Piracém	- Abaré - Anhanga - Jaci - Jurupari - Tupã	- Choça - Itaoca - Oca - Ocara - Taba	- Aracati - Corisco - Guarani - Irerê - Maracatim - Maranduba - Pocema - Pororoca - Sopé - Ticum

Fonte: A autora

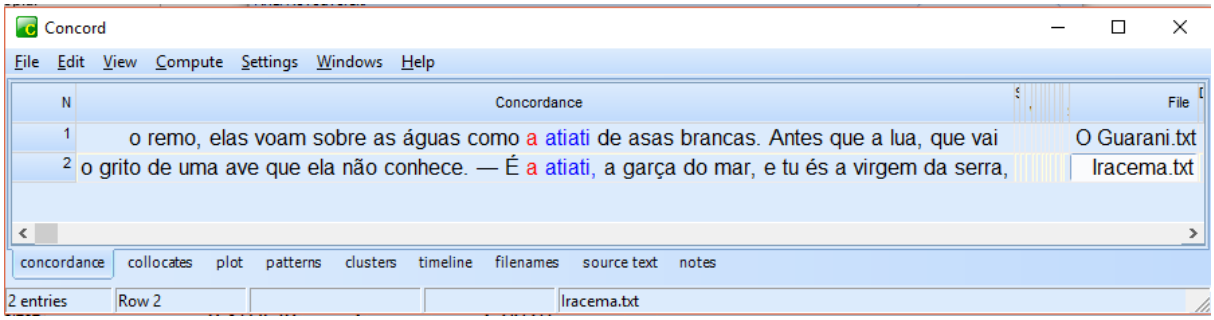
A seguir, buscamos esclarecer como as linhas de concordância obtidas por meio da ferramenta *Concord* são fundamentais para a elaboração do Vocabulário e das descrições dos vocábulos indianistas de Alencar.

4.2.9 Extração e análise das linhas de concordâncias com o léxico indianista

Para a elaboração das descrições dos vocábulos indianistas e inserção no Vocabulário, utilizamos a ferramenta *Concord* do *WST*. Essa ferramenta permite o acesso aos seus contextos linguísticos de ocorrência; dessa maneira, os traços semântico-contextuais podem ser recuperados para as descrições. Além disso, as linhas de concordâncias nos forneceram os contextos abonatórios, que nos auxiliaram nas descrições para o vocabulário.

A visualização das linhas de concordâncias possibilita observar a palavra candidata a vocábulo e o entorno textual em que é usada. Por exemplo, a Figura 24 mostra as linhas de concordância da palavra **atiati** e, pelo contexto abonatório, já é possível abstrair o significado pretendido pelo autor por meio da sua própria explicação: "É a <atiati>, a garça do mar, e tu és a virgem da serra, que nunca desceu às alvas praias onde arrebentam as vagas". (ALENCAR, 1965, p. 77). Essa explicação do autor é confirmada pela definição apresentada pelo dicionário GD "**Atyaty** – - s. Gaivota" (p. 391). Da mesma maneira, em outros casos, Alencar propõe uma explicação do vocábulo empregado no próprio corpo do texto, em aposto, porém não explicita a etimologia.

Figura 24: Listas das linhas de concordância do vocábulo **atiati**



Fonte: A autora, por meio da ferramenta *Concord* do *WST*.

Para ratificar a importância das linhas de concordâncias para extração da descrição dos vocábulos, por meio dos contextos, apresentamos o vocábulo **jatobá**, com suas linhas de concordâncias, como apresentamos na Figura 25.

Figura 25: Lista das linhas de concordâncias do vocábulo **jatobá**

Fonte: A autora, por meio da ferramenta *Concord* do *WST*.

É possível notar que o vocábulo é utilizado nos romances *Iracema* e *Ubirajara* com duas acepções. As linhas 1, 4, 6, 12 e 13 de concordância o vocábulo é grafado em letra minúscula e se refere a uma espécie de árvore; já nas demais linhas, o vocábulo é utilizado para nomear um personagem do romance *Iracema*, portanto um antropônimo. Ressaltamos

que a ferramenta *Concord* foi essencial para a percepção das duas acepções para o vocábulo **jatobá** utilizadas pelo autor José de Alencar.

Encerramos esta seção, que tratou sobre a importância das linhas de concordância para averiguar o contexto abonatório do vocábulo na obra, para explicitarmos sobre como os dicionários de exclusão e o *corpus* de consulta foram utilizados na descrição etimológica dos vocábulos de Alencar.

4.2.10 Utilização de dicionários e de **corpus** *de consulta, na análise e descrição etimológica do léxico de Alencar*

O léxico é o conjunto de vocábulos de uma língua, considerado um sistema aberto, portanto não é um produto estático, mas dinâmico no sentido de modificar-se constantemente em razão do uso que os falantes fazem dos itens lexicais dessa língua. Assim, conforme Biderman (2001), a inventividade humana cria novas significações para os vocábulos já existentes, assim como cria novos vocábulos de acordo com as necessidades do momento.

Em razão disso, os dicionários de língua, especialmente os mais antigos, nos oferecem uma parcela de informação sobre o léxico de uma língua em determinado momento da história. Portanto, para pesquisar sobre a etimologia dos vocábulos em Alencar, utilizamos quatro dicionários anteriores às publicações de suas obras que compõem o *corpus* de estudo de nossa pesquisa: um, exclusivamente de língua indígena, *Dicionário de língua Tupy*, de A. Gonçalves Dias (1858); e três de Língua Portuguesa: *Dicionário da Lingua Brasileira*, de Luiz Maria da Silva Pinto (1832); *Diccionário da Lingua Portugueza*, de Antonio de Moraes Silva (1789) e *Vocabulario Portuguez e Latino*, de Raphael Bluteau (1712).

A escolha das obras lexicográficas mencionadas se deveu ao fato de elas serem escritas em Língua Portuguesa, serem dicionários monolíngues e pela possibilidade de acesso. Apresentamos algumas características dos dicionários:

1. **Vocabulário Portugues e Latino**, de Raphael Bluteau (1712) - é considerado um marco dos estudos léxicos da Língua Portuguesa.

2. **Diccionnario da Língua Portuguesa**, de Antonio de Moraes Silva (1789) - segundo o próprio autor, foi elaborado de maneira mais enxuta que o Bluteau, o que proporcionou ideias mais claras e exatas.
3. **Dicionário da Lingua Brasileira**, de Luiz Maria da Silva Pinto (1832) – é um dicionário com 1132 páginas, no qual o autor se propõe a colocar todos os Vocábulos ao seu alcance e reforça que não se trata somente de termos indígenas, como se presumiria, mas de língua portuguesa. Portanto ele é considerado um dicionário de língua portuguesa brasileira.
4. **Dicionário de língua Tupy**, de A. Gonçalves Dias (1858) - de acordo com a apresentação do próprio autor, esse dicionário foi organizado com a intenção de guardar a memória do povo indígena, por solicitação do Instituto Histórico e Geográfico Brasileiro. Foi um trabalho realizado com base nos estudos da língua indígena, preocupados com o desaparecimento dessa língua, da qual nos restam muitos vestígios.

A consulta a esses dicionários nos propiciou informações sobre a data aproximada da utilização do vocábulo na língua portuguesa. Ressalvamos que nenhum dicionário, abarca todo o léxico de uma língua, porém ele pode ser também o ponto de partida para uma análise lexical. A iniciativa de utilizar quatro dicionários é uma tentativa de se identificar o momento de utilização de um determinado vocábulo. Entendemos também que generalizações, em relação ao emprego ou à existência de um determinado item lexical, não poderão ser feitas com base apenas nos dicionários consultados. Dessa maneira, nossas interpretações relacionadas à análise etimológica dialogam com os quatro dicionários apresentados sem a pretensão de esgotamento das possibilidades de interpretação de qualquer item lexical.

Para análise do léxico indianista que selecionamos em nosso *corpus* recorremos, prioritariamente, a fontes que podem nos assegurar uma etimologia mais próxima do uso pelo autor. Assim nos baseamos nos dicionários e no *corpus* mencionados, em Schwamborn (1998), uma estudiosa que empreendeu robustas pesquisas sobre Alencar, assim como consideramos as notas escritas pelo próprio Alencar ao final das obras e no contexto de uso no corpo das narrativas dos romances. Consideramos as explicações que o autor apresenta para alguns vocábulos como o faz em “mas os tabajaras, seus inimigos, por escárnio os apelidavam potiguaras, comedores de camarão”, neste caso é possível abstrair parte do significado de **potiguaras**, com base no texto da narrativa. Recorremos a uma ou a todas as fontes, quando

necessário e pertinente para a identificação da etimologia dos vocábulos; portanto, se não houver necessidade de recorrer aos dicionários, eles não são citados. Esse esclarecimento é importante para justificar os casos em que os dicionários não foram mencionados.

Para a análise relacionada à etimologia dos vocábulos considerados indígenas, de acordo com os critérios mencionados, adotamos apenas os quatro dicionários explicitados nesta seção, por serem anteriores à publicação das obras de Alencar, porém na elaboração das fichas lexicográficas, outros dicionários, contemporâneos a Alencar, também foram utilizados.

Concluímos esta seção, para explicitarmos os procedimentos de elaboração das fichas lexicográficas.

4.2.11 Elaboração das fichas lexicográficas com o léxico indianista identificado e analisado

Os dicionários podem ser impressos, eletrônicos ou disponibilizados *online*. Os eletrônicos, caracterizamos aqueles distribuídos em CDs, que podem ser instalados nos computadores. Já os disponibilizados *online* são aqueles cuja pesquisa é livre e dependem somente de conexão com *internet*.

Embora nos propomos a organizar um vocabulário *online*, cujas características se assemelham aos dicionários eletrônicos, portanto, sem uma forma fixa para a consulta, organizamos uma ficha, constituída pelos elementos que julgamos procedentes para nosso trabalho. Como também não há critérios rígidos e únicos para sistematizar o trabalho dos lexicógrafos, elaboramos um modelo de ficha baseada nos procedimentos de produção dos verbetes de Ávila (2004); Ávila e Martins (2008) e Schreiner (2012). As fichas nos fornecem os dados que serão inseridos no vocabulário *online*, para que o consulente possa optar pelo tipo de pesquisa. Os detalhes sobre as possibilidades de busca serão pormenorizados no Capítulo 06 desta tese.

As fichas lexicográficas estão constituídas com as seguintes informações:

- **Entrada** – é o objeto da descrição, ou seja, trata-se do vocábulo ou item lexical que abre o verbete. Os vocábulos serão grafados preconizando a forma masculina singular, exceto os substantivos e adjetivos cuja marcação de gênero interfira no significado.

- **Classe gramatical** – é informada a classe gramatical, adotando como critério utilizar a forma abreviada: s.m. – substantivo masculino; s.f. – substantivo feminino; adj. – adjetivos;

- **Romance em que aparece** – é identificada a obra de Alencar em que aparece o vocábulo. Quando o vocábulo for utilizado em mais de um romance também é informado. Utilizamos os nomes das obras entre parênteses.

- **Campo semântico** – são mencionados os campos semânticos de acordo com a classificação em: antropônimos, topônimos, etnônimos, flora, fauna, utensílios e vestimentas, alimentos, espiritualidade e os diversos;

- **Descrição contextual** – apresentamos uma descrição elaborada a partir de informações retiradas dos romances, por meio dos contextos de abonação, e nas definições apresentadas pelos dicionários de exclusão.

- **Informações sobre a etimologia e o processo de formação** – quando possível, é informado o processo de formação e a etimologia do vocábulo;

- **Passagem abonatória** – em um campo específico, é informada, pelo menos uma passagem abonatória para o vocábulo. São transcritas quantas passagens foram necessárias para esclarecer a descrição, portanto não há uma padronização em relação à quantidade de abonações. Quando houve a ocorrência do vocábulo em mais de um romance, é também transcrita pelo menos uma passagem abonatória de cada um. Como não há consenso no que se refere à extensão das passagens abonatórias, é transcrita a extensão de texto que julgamos necessário e, ao mesmo tempo, suficiente para abstrair a descrição do vocábulo;

- O vocábulo em análise é destacado na passagem abonatória entre os sinais < >, por exemplo <zabelê> e a obra de onde se transcreveu a passagem abonatória está escrita entre parênteses ao final da transcrição, por exemplo “A virgem disse e desapareceu na selva. Os olhos de Jaguarê seguiram o passo ligeiro da formosa caçadora, como o guaxinim que rasteja a <zabelê>”. (UBIRAJARA)

- **Definição dos dicionários consultados** – são inseridas, neste item, as definições constantes de todos os dicionários consultados. Nos casos em que não se figurar em algum deles, utilizamos a forma n/c para indicar que o vocábulo não consta do dicionário em questão.

- Caso o vocábulo apareça com duas acepções, são criadas fichas diferentes. Isso porque se trata de definição contextual, o que podem ser atribuídas definições distintas para o mesmo vocábulo, de acordo com a acepção. Por exemplo, **jatobá** que é empregado como antropônimo e elemento da flora. Assim esse vocábulo tem duas fichas lexicológicas;

- **Notas** – serão elaboradas, quando necessário, notas para esclarecer quaisquer questões sobre os vocábulos inseridos;

A seguir o modelo da ficha lexicográfica elaborada.

Ficha Nº		
Vocábulo:	Classe gramatical:	Campo semântico:
Descrição contextual		
Etimologia e/ou Processo de formação		
Passagens abonatórias:		
Definição dos dicionários consultados: GD: TS: AGC: LCT:		
Nota:		

A ficha lexicográfica permite uma visualização mais ampla do vocábulo em estudo, por exemplo é possível analisar o tratamento que os dicionários, tanto os mais antigos como os mais atualizados, dão aos verbetes. Assim como também é possível visualizar o contexto de abonação de uma ou mais obras do autor. Além de analisar o vocábulo selecionado, a ficha também constitui uma ferramenta para composição do vocabulário que será disponibilizado *online*, assim como auxiliará no trabalho de comparação e quantificação dos dados. Na próxima seção, trataremos de questões relacionadas à elaboração de verbetes.

4.2.12 Elaboração de proposta de verbete

O verbete corresponde a cada uma das entradas de um dicionário, de um glossário ou um vocabulário, acompanhada das informações que o lexicógrafo escolhe dispor em seu trabalho. O verbete é considerado um repertório lexical em que são reunidos os dados relativos à unidade lexical e é composto de pelo menos dois elementos: entrada e enunciado. (BARROS, 2004). Há, porém, diferentes níveis de complexidade em relação à organização de um verbete.

No caso de nossa pesquisa, as unidades léxicas que constituem as entradas dos verbetes são os vocábulos indianistas dos romances que compõem o *corpus* de estudo. É fundamental destacar que apresentamos apenas exemplos de verbetes que servirão como modelo para elaboração dos demais, assim como servirão de base para organização do Vocabulário *online*.

Então, na amostragem do verbete, apresentamos a entrada, seguida das informações, de classe e gênero. Em seguida, apresentamos a etimologia e/ou processo de formação, o campo semântico a que o vocábulo pertence e a descrição. Finalmente, acrescentamos as passagens abonatórias que sustentam a descrição apresentada com o respectivo romance em que ocorre.

Jatobá. s.m. Etim. De *ya – atã – yba*. Elemento da flora.. Árvore da família das leguminosas que possui o fruto com uma casca dura no formato de uma grande vagem que se assemelha a um estojo, capaz de cair de grande altitudes e não se quebrar. Dentro do fruto, a polpa tem o formato semelhante a uma banana. O fruto da árvore também leva o nome de jatobá. Passagens abonatórias: 1)*"Poti levou o cristão aonde crescia um frondoso <jatobá>, que afrontava as árvores do mais alto píncaro da serrania, e quando batido pela rajada, parecia varrer o céu com a imensa copa." (Iracema)* 2)*"— Pais da sabedoria, abarés, olhai aquele <jatobá> que se levanta no meio da campina, e que eu só posso ver agora na sombra de minha alma." (Ubirajara)*

Como já mencionado, a elaboração dos verbetes são um passo prévio para a constituição do Vocabulário *online*. Por fim, vale ressaltar que, por se tratar de uma proposta de criação de um Vocabulário *online*, a forma de acesso será apresentada no Capítulo 7 "Léxico indianista de José de Alencar: proposta para um vocabulário *online"*.

No capítulo seguinte, analisaremos os dados obtidos a partir do contraste do autor com ele mesmo e do autor com os outros *corpora* de referência, além de uma análise baseada também na pesquisa com o *corpus* de consulta de Davies (2016).

CAPÍTULO 5 - ANÁLISE CONTRASTIVA DO *CORPUS* DE ESTUDO DE ALENCAR COM OS *CORPORA* DE REFERÊNCIA

Este capítulo parte do intuito de investigar se José de Alencar utiliza um léxico específico em suas obras indianistas. Para isso, contrastamos o léxico das obras indianistas de Alencar, *corpus* de estudo, com as demais não indianistas, o CorpRef-Alencar. Além disso, o objetivo também consiste em certificar se esse léxico resultante das obras indianistas apresentaria modificação, por meio da análise das palavras-chave em contraste com outros *corpora* de referência de extensão muito maior do que o *corpus* de estudo.

Em relação à constituição de um *corpus* de referência para extração de palavras-chave por meio do *software*, Berber Sardinha (2004, p. 100) estabelece que devem ser obedecidos os seguintes princípios: primeiro, que o "*corpus* de referência não deve conter o *corpus* de estudo, pelo menos não deliberadamente e por completo", isso porque, segundo o autor, a comparação entre o *corpus* de estudo e o *corpus* de referência do mesmo tipo tende a filtrar as palavras-chave, já que elas também aparecerão no *corpus* de referência. Por outro lado, quando o *corpus* de referência pertence a gêneros diferentes do *corpus* de estudo, a tendência a excluir palavras consideradas genéricas é menor. Por isso, segundo Berber Sardinha (2004), como segundo princípio, estabelece que o *corpus* de referência deve incluir vários gêneros, para que as palavras-chave não percam a evidência.

Apesar do princípio estabelecido pelo autor, de que o *corpus* de referência deve pertencer a gêneros diferentes do *corpus* estudado, ele afirma também que "o *corpus* deve ser adequado aos interesses do pesquisador, que deve ter uma questão a investigar para a qual necessite de um *corpus* específico" (BERBER SARDINHA, 2004, p. 29). Com vistas a atender ao objetivo desta tese, proposto neste capítulo, investigamos se Alencar utiliza um léxico característico nas obras indianista, por meio do contraste do léxico com as demais obras do mesmo autor. Esta parte da pesquisa visa, portanto, ao contrastaste do autor com ele mesmo. Não é um procedimento recorrente entre os trabalhos já desenvolvidos, o contraste de um autor com ele mesmo no âmbito de sua produção literária ou científica.

Após a observância desses critérios, procedemos à extração das listas de palavras-chave de cada um dos *corpora*, seguidos da comparação estatística, a partir da utilização das ferramentas *KeyWords* e *WordList* do *WST*, cujos resultados serão apresentados nas seções seguintes.

5.1 Análise das palavras-chave do *corpus* de estudo

Toda palavra de um *corpus* tem um potencial para análise. Portanto, para aguçar nosso estudo sobre o léxico indianista de Alencar, usamos a ferramenta *KeyWord* do *WST* para investigarmos as palavras características do *corpus* de estudo e propomos uma análise. A Figura 26 mostra a tela do *KeyWords* com as 20 palavras mais chave, organizadas pelo valor da chavicidade (*Keyness*) do *corpus* de estudo em relação ao *corpus* de referência de Alencar.

Figura 26: Palavras-chave do *corpus* de estudo em contraste com o *corpus* de referência CorpRef-Alencar

Corpus de Estudo _ Corpus de Referência Alencar.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	RC.	RC. %	Keyness	P	Lemm
1	PERI	727	0,48	1		3.266,57	0,0000000000	
2	CECÍLIA	458	0,30	1		2.052,55	0,0000000000	
3	GUERREIRO	458	0,30	49		1.755,31	0,0000000000	
4	ÍNDIO	338	0,22	50		1.237,78	0,0000000000	
5	GUERREIROS	297	0,20	23		1.179,51	0,0000000000	
6	ÁLVARO	279	0,19	33		1.055,26	0,0000000000	
7	IRACEMA	220	0,15	3		961,22	0,0000000000	
8	LOREDANO	174	0,12	0		784,83	0,0000000000	
9	CHEFE	260	0,17	97		776,67	0,0000000000	
10	UBIRAJARA	169	0,11	0		762,27	0,0000000000	
11	ANTÔNIO	300	0,20	215	0,02	701,09	0,0000000000	
12	ITALIANO	173	0,11	12		694,12	0,0000000000	
13	AVENTUREIROS	171	0,11	21		643,38	0,0000000000	
14	CABANA	220	0,15	96		625,54	0,0000000000	
15	MARIZ	175	0,12	39		594,77	0,0000000000	
16	ITAQUÊ	114	0,08	0		514,16	0,0000000000	
17	VIRGEM	214	0,14	151	0,01	503,65	0,0000000000	
18	POTI	102	0,07	0		460,03	0,0000000000	
19	TABA	105	0,07	8		417,54	0,0000000000	
20	ARACI	97	0,06	3		411,19	0,0000000000	

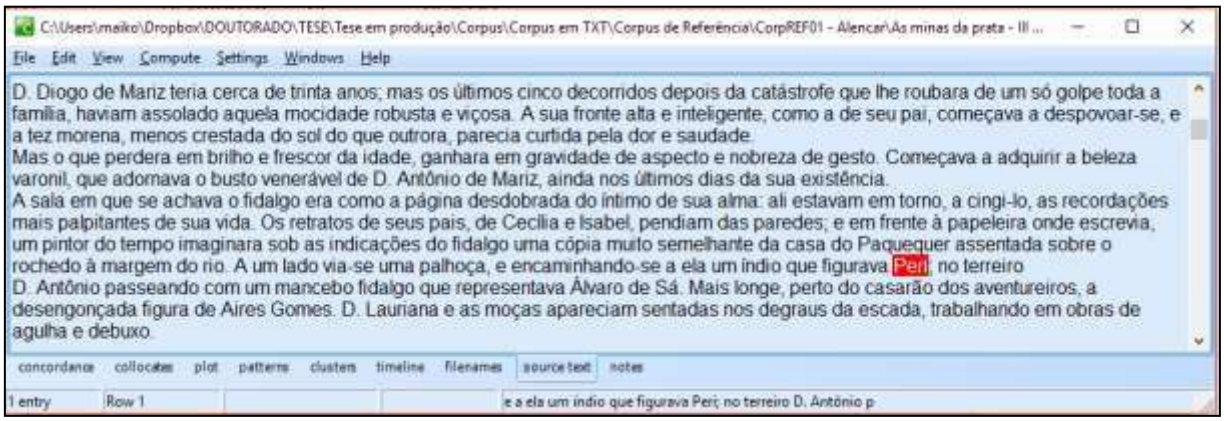
KWs plot links clusters filenames source text notes

Fonte: A autora, por meio do *WST*

Analisamos, inicialmente, a lista de palavras por ordem de chavicidade, ou seja, em ordem decrescente, como apresentado na coluna *Keyness*. Neste caso, a palavra “mais

chave" do *corpus* de estudo é **Peri**, porque possui um valor de chavicidade de 3.266,57, o maior da lista. Essa palavra tem apenas uma ocorrência no *corpus* de referência e foi utilizada na obra *As Minas de Prata*. Alencar, em *As Minas de Prata*, propõe uma intersecção entre os personagens dessa obra com os do livro *O Guarani*, conforme o trecho extraído, por meio da ferramenta *Concord* do *WST,* mostrado pela Figura 27.

Figura 27: Vista do trecho do romance *As Minas de Prata* em que Alencar emprega o vocábulo **Peri**



Fonte: A autora, por meio da ferramenta *Concord* do *WST*

As Minas de Prata é considerado um romance histórico e foi publicado sete anos depois de *O Guarani.* Na primeira publicação, Alencar intitulou o romance com o subtítulo *Continuação do Guarani,* porém este subtítulo foi extraído na edição de 1865/66. Inicialmente, a ação seria a continuação de *O Guarani*, pois um dos personagens de *As Minas de Prata*, o personagem D. Diogo de Mariz é filho de D. Antonio de Mariz e, portanto, irmão de **Ceci**, a heroína de *O Guarani*. D. Antonio era o proprietário do solar onde se passou a história de amor do índio **Peri** por **Ceci**.

No romance *As Minas de Prata*, D. Diogo é portador do roteiro que levaria às minas de prata, porém esse roteiro é de propriedade de outro personagem do romance, Estácio. A história continua e marca a presença de um personagem que transitou de um romance para o outro, porém, em razão do próprio roteiro da história escrita por Alencar, ele não participa das ações de Estácio em busca das minas. (MARCO, 1993; VASCONCELOS, 2011). Não nos ateremos à descrições e análises do romance *As Minas de Prata*, porém, esse breve comentário e a apresentação do trecho deste texto que compõe o *corpus* de referência, Figura

27, demonstra a frequência um e justifica a utilização desse vocábulo por Alencar em obras não indianistas.

O emprego do vocábulo **Peri** em *As Minas de Prata*, destaca que, mesmo sendo utilizado no *corpus* não indianista, Alencar o emprega, retomando o personagem de *O Guarani*. Como é possível observar, Alencar descreve a sala em que D. Diogo está, em ênfase no detalhe do quadro da parede onde o índio **Peri** também é retratado na pintura que orna o ambiente, conforme o trecho

> Os retratos de seus pais, de Cecília e Isabel, pendiam das paredes; em frente à papeleira onde escrevia, um pintor do tempo imaginara sob as indicações do fidalgo uma cópia muito semelhante da casa do Paquequer assentada sobre o rochedo à margem do rio. A um lado via-se uma palhoça, e encaminhando-se a ela um índio que figurava Peri. (ALENCAR, s/p.).

Dessa forma, o emprego do vocábulo é, marcadamente, associado ao personagem índio do romance *O Guarani*, tanto no *corpus* de estudo quanto no de referência. Dessa maneira, é um vocábulo empregado, especificamente, por Alencar com o propósito indianista.

No que se refere à chavicidade, o valor representa a extensão da diferença das porcentagens decorrentes das frequências de **Peri** nos dois *corpora*. No *corpus* de estudo, as ocorrências da palavra **Peri** (727) correspondem a 0,48% de ocorrências do *corpus* em totalidade. Em contrapartida, no *corpus* de referência, a ocorrência não foi mensurada em termos de porcentagem pela insignificância da ocorrência, que é uma. Destacamos, assim, a diferença entre as ocorrências, de 727 por 1, em números absolutos de ocorrências.

É possível também visualizar, dentre as 20 primeiras palavras mais chave do *corpus* de estudo, vocábulos de frequência zero no *corpus* de referência, como **Loredano**, **Ubirajara**, **Itaquê** e **Poti**. A frequência zero se justifica, pois essas palavras são nomes dos personagens dos livros que compõem o *corpus* de estudo: **Itaquê** e **Ubirajara** do livro *Ubirajara*; **Poti** de *Iracema*; e **Loredano** de *O Guarani*.

Em observância ao princípio estabelecido por Berber Sardinha (2004, 2009), de que o *corpus* de referência deve ser composto por gêneros diferentes do *corpus* de estudo e visando atender ao objetivo de tese, geramos também as *KeyWords* com os outros *corpora* de referência: CorpRef-Lácio-Web, CorpRef-AcadTeses e CorpRef-Nov e, assim, proceder à análise do contraste.

Partimos do princípio de que o *corpus* de referência é "também conhecido como '*corpus* de controle' e funciona como termo de comparação para análise". (BERBER SARDINHA, 2009, p. 194). Ainda segundo o autor, a função do *corpus* de referência é estabelecer uma norma com que se fará a comparação das frequências das palavras-chave do *corpus* de estudo. Assim sendo, foram geradas mais três listas de palavras-chave para a comparação entre os quatro *corpora* de referência. A Figura 28 apresenta as palavras-chave do *corpus* de estudo em contraste com o *corpus* CorpRef-AcadTeses; a Figura 29 o contraste com o *corpus* CorpRef-Lácio-Web; e a Figura 30 com o *corpus* CorpRef-Nov.

Figura 28: Contraste com o *corpus* de referência CorpRef-AcadTeses

N	Key word	Freq.	%	RC.	RC. F	Keyness
1	PERI	727	0,48	646		7.619,98
2	GUERREIRO	458	0,30	868		4.287,73
3	CECÍLIA	458	0,30	897		4.263,56
4	GUERREIROS	297	0,20	605		2.745,62
5	OLHOS	375	0,25	4.421		2.289,40
6	ÁLVARO	279	0,19	1.134		2.250,56
7	LOREDANO	174	0,12	3		2.246,69
8	IRACEMA	220	0,15	288		2.184,87
9	ÍNDIO	338	0,22	3.772		2.098,39
10	UBIRAJARA	169	0,11	30		2.042,98
11	MARIZ	175	0,12	56		2.034,46
12	CABANA	220	0,15	595		1.930,25
13	FIDALGO	193	0,13	271		1.896,48
14	VIRGEM	214	0,14	1.100		1.635,94
15	AVENTUREIR(	171	0,11	332		1.593,93
16	MENINA	215	0,14	1.479		1.528,87
17	ITAQUÊ	114	0,08	0		1.491,84
18	ISABEL	196	0,13	1.881		1.272,21
19	TABA	105	0,07	29		1.234,16
20	POTI	102	0,07	25		1.208,88

Fonte: A autora, com base na ferramenta *KeyWords*

Figura 29: Contraste com o *corpus* de referência CorpRef-Lácio-Web

N	Key word	Freq.	%	RC. Freq.	RC.	Keyness
1	PERI	727	0,48	1		5.649,34
2	GUERREIRO	458	0,30	97		3.057,37
3	CECÍLIA	458	0,30	146		2.905,64
4	ÍNDIO	338	0,22	123		2.102,93
5	GUERREIROS	297	0,20	42		2.061,00
6	ÁLVARO	279	0,19	75		1.810,49
7	OLHOS	375	0,25	496		1.750,75
8	IRACEMA	220	0,15	5		1.665,67
9	CABANA	220	0,15	10		1.631,56
10	FIDALGO	193	0,13	5		1.456,65
11	ANTÔNIO	300	0,20	388		1.410,10
12	VIRGEM	214	0,14	58		1.387,16
13	LOREDANO	174	0,12	0		1.355,11
14	MARIZ	175	0,12	10		1.285,50
15	AVENTUREIROS	171	0,11	6		1.279,58
16	ISABEL	196	0,13	52		1.273,90
17	SENHORA	235	0,16	185		1.261,59
18	MENINA	215	0,14	157		1.174,30
19	CHEFE	260	0,17	405		1.151,65
20	UBIRAJARA	169	0,11	58		1.060,54

Fonte: A autora, com base na ferramenta *KeyWords*

Figura 30: Contraste com o *corpus* de referência CorpRef-Nov

N	Key word	Freq.	%	RC.	RC. %	Keyness
1	PERI	727	0,48	0		2.142,68
2	GUERREIRO	458	0,30	3		1.314,60
3	CECÍLIA	458	0,30	4		1.305,35
4	GUERREIROS	297	0,20	0		874,69
5	ÍNDIO	338	0,22	26		821,71
6	ÁLVARO	279	0,19	9		746,24
7	CABANA	220	0,15	0		647,83
8	IRACEMA	220	0,15	5		602,48
9	ISABEL	196	0,13	3		547,58
10	ANTÔNIO	300	0,20	76	0,02	544,58
11	VIRGEM	214	0,14	14		532,20
12	LOREDANO	174	0,12	0		512,33
13	MARIZ	175	0,12	2		494,41
14	FIDALGO	193	0,13	10		493,80
15	AVENTUREIROS	171	0,11	1		491,73
16	UBIRAJARA	169	0,11	5		454,86
17	CHEFE	260	0,17	85	0,02	424,69
18	ITALIANO	173	0,11	14		417,18
19	OLHOS	375	0,25	234	0,05	415,03
20	SENHORA	235	0,16	72	0,01	395,08

Fonte: A autora, por meio da ferramenta *KeyWords*

Ao procedermos à comparação entre as quatro listas de palavras-chave, notamos uma certa regularidade, pois 13 das vinte primeiras palavras são chave no *corpus* de estudo em relação aos quatro *corpora* de referência: CorpRef-Alencar, CorpRef-Nov; CorpRef-Lácio-Web; e CorpRef-AcadTeses. O fato de o *corpus* de referência, CorpRef-Alencar, ser composto por obras de Alencar, assim como o *corpus* de estudo, não invalidou a nossa escolha. Pelo contrário, confirmou a hipótese de que Alencar utiliza um léxico específico em suas obras indianistas. Isso pode ser corroborado, pelo fato de que, no conjunto, as palavras-chave extraídas por meio do contraste de Alencar com ele mesmo também foram identificadas

no contraste com os demais *corpora* de referência. Além disso, outras seis palavras são, igualmente, chave em três dos *corpora* de referência; outras cinco são chave em dois dos *corpora* de referência e apenas a palavra **Araci** aparece como palavra-chave em contraste com um único *corpus* de referência que é o *corpus* CorpRef-Alencar[36].

Merece observação, também, o fato de que **Peri** aparece com 646 ocorrências no CorpRef-AcadTeses contra 726 do *corpus* de estudo, entretanto, isso não impediu que fosse a palavra mais chave do estudo, uma vez que a diferença de extensão entre os *corpora* é de 642 vezes maior.

Berber Sardinha afirma que

> o conteúdo do *corpus* de referência influencia quais palavras-chave serão identificadas, de tal modo que é possível antecipar o tipo de influência que um *corpus* de referência terá no resultado da lista de palavras-chave se compararmos os perfis dos corpora, com relação a aspectos como: os gêneros incluídos, os assuntos, os períodos, a autoria etc. (BERBER SARDINHA, 2009, p. 194).

O conteúdo do *corpus* de referência pode influenciar na lista das palavras-chave, porém, no caso desta pesquisa, esta influência não refletiu nos contrastes realizados entre o *corpus* de estudo e os *corpora* de referência, pois, mesmo os *corpora* tendo extensões diversas: CorpRef-Alencar com 61.121 formas; o CorpRef-AcadTeses com 620.068; o CorpRef-Lácio-Web com 130.020 e o *corpus* CorpRef-Nov com 43.119 formas, as palavras-chave com maior chavicidade foram atestadas nos diferentes contrastes, o que comprova a especificidade do léxico nas obras indianistas de Alencar. Outro fator que corrobora para essa afirmação é que, apesar de os *corpora* também apresentarem perfis diferenciados, conforme descrito no capítulo de Metodologia, as palavras-chave, entre as primeiras da lista, se repetiram.

Cabe destacar que, apesar de o *corpus* de referência de Alencar ser formado também por romances do autor, a maioria das palavras-chave estão repetidas, isso nos leva a ratificar a afirmação, com base nos dados apresentados, de que José de Alencar utiliza um léxico específico em suas obras indianistas. Nesse sentido, o aspecto dos gêneros presente nos *corpora*, seja de estudo, seja no de referência, não foi determinante para modificar os resultados nas listas de palavras-chave.

[36] Pode-se certificar no quadro do Apêndice A.

Outro fator relevante é o fato de que, como afirma Berber Sardinha (2009), é possível identificar a temática de um *corpus*, como também traçar o perfil lexical de um autor por meio das palavras-chave desse *corpus*. Isso é possível perceber em Alencar que utiliza palavras que nos remetem ao universo dos indígenas como **taba**, **virgem**, **cabana**, **aventureiros**, **chefe, guerreiros**, **guerreiro** e a própria palavra **índio**. O autor também utiliza nomes próprios que são palavras-chave nas listas geradas, como: **Peri**, **Cecília**, **Álvaro**, **Iracema**, **Loredano**, **Ubirajara**, **Antonio**, **Mariz, Itaquê**, **Poti**, **Araci** e **Isabel,** que **s**ão palavras escolhidas para nomear seus personagens.

Durante os procedimentos de extração das palavras-chave, de observação atenta e análise relacionada às frequências das palavras no contraste do *corpus* de estudo com os *corpora* de referência, percebemos que havia palavras cuja frequência é zero no *corpus* de referência, isto é, palavras que ocorreram mesmo com frequência baixa no *corpus* de estudo, no *corpus* de referência não registram nenhuma ocorrência. Assim, o *corpus* nos guiou para a análise também dessas palavras de frequência zero e, em razão disso, decidimos adotar a perspectiva de análise guiada pelo *corpus*, conforme Berber Sardinha (2004; 2009).

O *corpus* nos levou a descobrir outros aspectos não programados anteriormente, porém com o manuseio e análise dos dados, a medida que foram tratados, formulamos novas hipóteses. Essa análise tornou-se relevante, na medida em que essas palavras, as de frequência zero, poderiam revelar um acervo lexical específico, utilizado pelo autor, para compor seus romances indianistas. Por esta razão, empreitamos análise também das palavras de frequência zero, o que, no princípio da pesquisa, não seria, especificamente, objeto de estudo.

Apresentamos, portanto, na seção seguinte, uma parcial das palavras-chave com frequência zero no contraste com o *corpus* de estudo e demais *corpora* de referência.

5.2 *Corpus* de estudo em contraste com os *corpora* de referência: análise das palavras de frequência zero

Com a atenção voltada para as palavras de frequência zero, obtivemos a lista de palavras de ocorrência zero por meio da ferramenta *KeyWords* do *WST*. Esta frequência é obtida por meio do ajuste da ferramenta *KeyWords* clicando na aba *RC.Freq*, cuja função é organizar a lista de palavras pela frequência no *corpus* de referência. A Figura 31 apresenta o

contraste com o *corpus* CorpRef-Alencar; a Figura 32, com o *corpus* CorpRef-AcadTeses; a Figura 33 com o *corpus* CorpRef-Lácio-Web; e a Figura 34 com o *corpus* CorpRef-Nov.

Figura 31: *Corpus* de estudo em contraste com *corpus* de referência CorpRef-Alencar

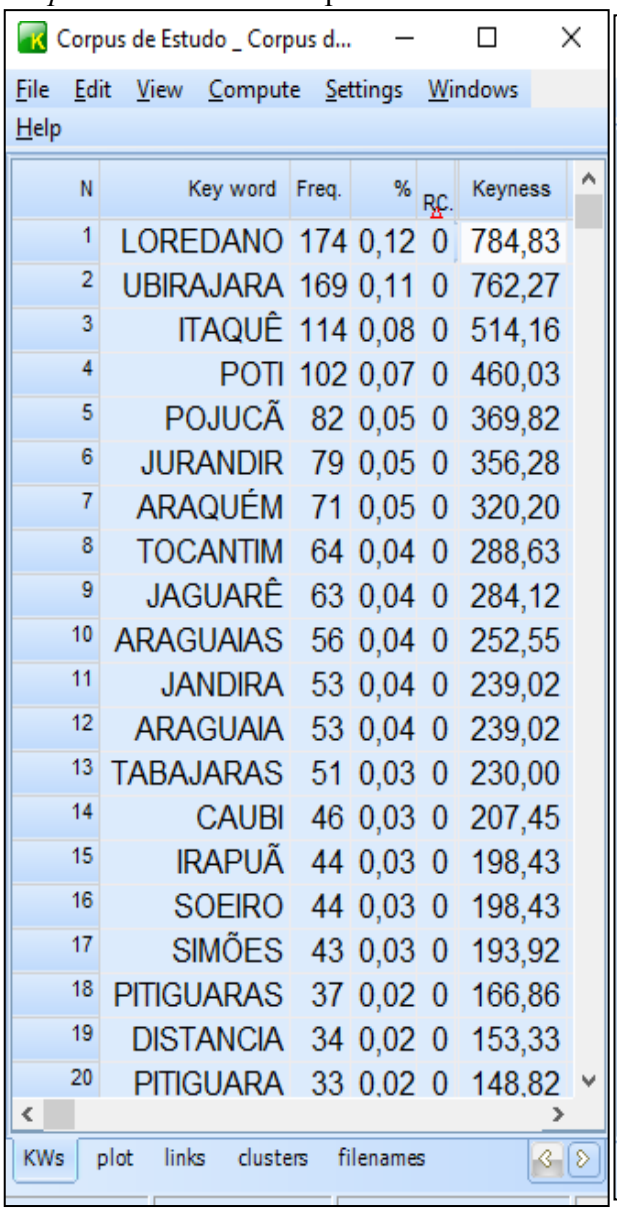
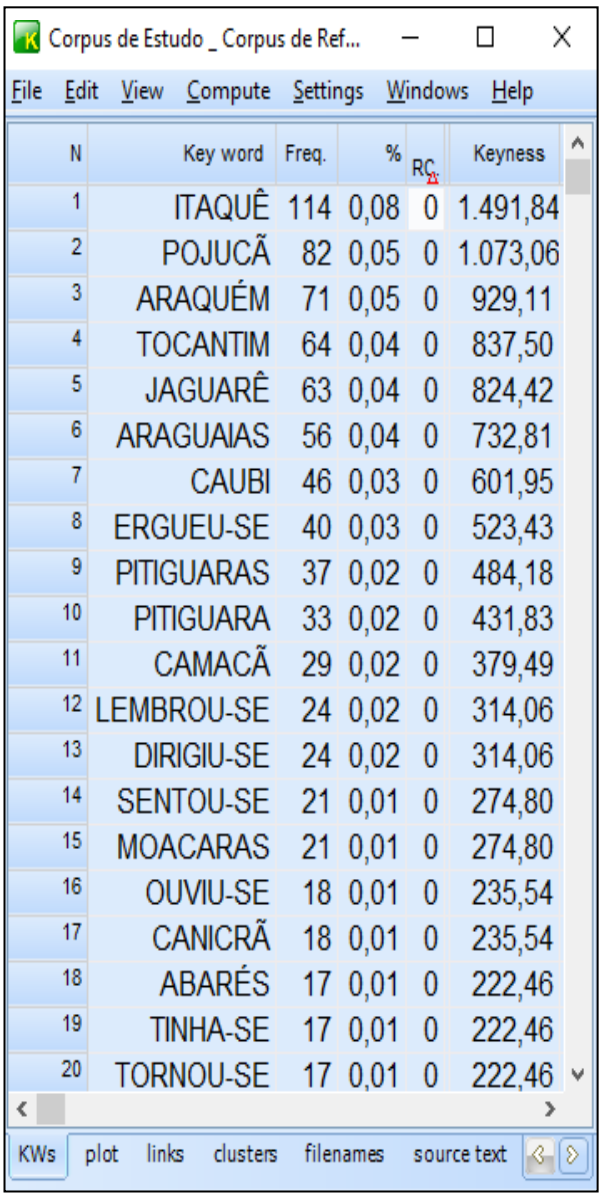
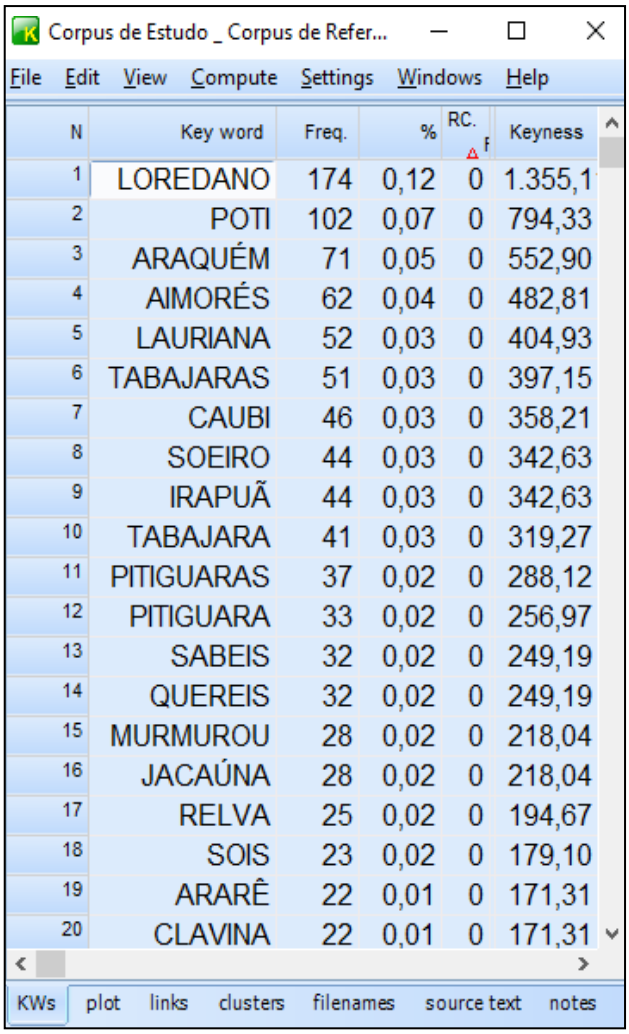


Figura 32: *Corpus* de estudo em contraste com o *corpus* de referência CorpRef-AcadTeses



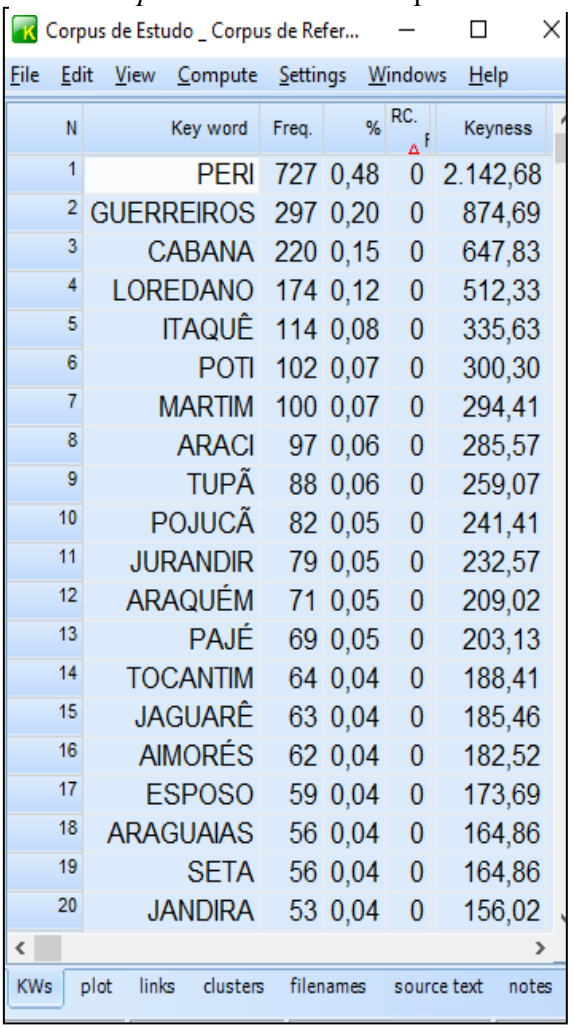
Fonte: A autora, por meio da ferramenta *KeyWords*

Figura 33: *Corpus* de estudo em contraste com o *corpus* de referência CorpRef-Lácio-Web



N	Key word	Freq.	%	RC.	Keyness
1	LOREDANO	174	0,12	0	1.355,1
2	POTI	102	0,07	0	794,33
3	ARAQUÉM	71	0,05	0	552,90
4	AIMORÉS	62	0,04	0	482,81
5	LAURIANA	52	0,03	0	404,93
6	TABAJARAS	51	0,03	0	397,15
7	CAUBI	46	0,03	0	358,21
8	SOEIRO	44	0,03	0	342,63
9	IRAPUÃ	44	0,03	0	342,63
10	TABAJARA	41	0,03	0	319,27
11	PITIGUARAS	37	0,02	0	288,12
12	PITIGUARA	33	0,02	0	256,97
13	SABEIS	32	0,02	0	249,19
14	QUEREIS	32	0,02	0	249,19
15	MURMUROU	28	0,02	0	218,04
16	JACAÚNA	28	0,02	0	218,04
17	RELVA	25	0,02	0	194,67
18	SOIS	23	0,02	0	179,10
19	ARARÊ	22	0,01	0	171,31
20	CLAVINA	22	0,01	0	171,31

Figura 34: *Corpus* de estudo em contraste com o *corpus* de referência CorpRef-Nov



N	Key word	Freq.	%	RC.	Keyness
1	PERI	727	0,48	0	2.142,68
2	GUERREIROS	297	0,20	0	874,69
3	CABANA	220	0,15	0	647,83
4	LOREDANO	174	0,12	0	512,33
5	ITAQUÊ	114	0,08	0	335,63
6	POTI	102	0,07	0	300,30
7	MARTIM	100	0,07	0	294,41
8	ARACI	97	0,06	0	285,57
9	TUPÃ	88	0,06	0	259,07
10	POJUCÃ	82	0,05	0	241,41
11	JURANDIR	79	0,05	0	232,57
12	ARAQUÉM	71	0,05	0	209,02
13	PAJÉ	69	0,05	0	203,13
14	TOCANTIM	64	0,04	0	188,41
15	JAGUARÊ	63	0,04	0	185,46
16	AIMORÉS	62	0,04	0	182,52
17	ESPOSO	59	0,04	0	173,69
18	ARAGUAIAS	56	0,04	0	164,86
19	SETA	56	0,04	0	164,86
20	JANDIRA	53	0,04	0	156,02

Fonte: A autora, por meio da ferramenta *KeyWords*

A primeira coluna traz a lista de palavras-chave (*KeyWords*) depreendidas pela confrontação entre o *corpus* de estudo e os *corpora* de referência. As palavras-chave são apresentadas em ordem decrescente de chavicidade, ou seja, as primeiras palavras, **Loredano**, **Peri** e **Itaquê** são as mais frequentes dentre aquelas de frequência zero nos *corpora* de referência. A segunda coluna registra a frequência do item no *corpus* de estudo; a terceira, a porcentagem do item em relação a todo o *corpus* de estudo. Já na coluna seguinte, em que aparecem os zeros, a ratificação de que estes itens não se encontram nos *corpora* de referência. A coluna que representa a porcentagem correspondente à representatividade do item em relação ao *corpus* de referência está em branco, em razão do baixo percentual.

Diferentemente das listas de palavras-chave geradas para obtenção das *KeyWords* contrastando o *corpus* de estudo com os *corpora* de referência, como já demonstrado, em que houve uma regularidade entre as 20 primeiras palavras-chave, no caso do contraste tomando por base as palavras de frequência zero apresenta uma diferença. Considerando a frequência zero e o critério chavicidade, apenas a palavra **Araquém** está entre as vinte primeiras em todos os *corpora* de referência. Porém, ainda assim, há nove palavras que constam de três dos *corpora*; cinco palavras constam entre as 20 de dois *corpora*; e 30 palavras estão entre as 20 mais frequentes alternando-se entre os *corpora*[37].

Observando ainda as palavras com frequência zero nos *corpora* de referências, verificamos que, no CorpRef-Alencar, do total dos 291 itens, 55 palavras são utilizadas apenas no *corpus* de estudo. Seguindo a mesma análise, obtivemos no CorpRef-AcadTeses dos 1.879 itens, 306 palavras com frequência zero no *corpus* de referência; já no CorpRef-Lácio-Web, dos 1.549 itens, 334 palavras do *corpus* de estudo têm frequência zero em relação ao de referência; por fim, no CorpRef-Nov, dentre os 477 itens, 104 palavras também têm frequência zero. O Quadro 8 organiza os números mencionados, obtidos por meio das *KeyWords* demonstrado nas Figuras 31 a 34.

Quadro 8: Quantidade de palavras com frequência zero em relação ao total de itens nos *corpora* de referência

Corpora de Referência	Quantidade total de itens obtidos a partir das *KeyWords*	Quantidade de palavras com frequência zero nos *corpora* de referência
CorpRef-Alencar	291 itens	55 palavras
CorpRef-AcadTeses	1.879 itens	306 palavras
CorpRef-Lácio-Web	1.549 itens	334 palavras
CorpRef-Nov	477 itens	104 palavras

Fonte: A autora

Por meio desses dados, é possível observar a produtividade de José de Alencar em suas obras indianistas. As palavras-chave colaboram, sobremaneira, para traçar o perfil léxico das obras indianistas de Alencar, uma vez que é relevante o número de palavras com frequência zero em relação ao número de itens de cada *corpus* de referência. Chamou-nos a

[37] O quadro do Apêndice B demonstra a frequência mencionada.

atenção, o fato de que alguns itens constantes entre os 20 primeiros da lista que, de primeira análise, tratava-se de nomes de personagens. Para constatar ou refutar essa primeira análise, consultamos todas as palavras nas obras do autor, por meio da ferramenta *Concord*, com a qual foi possível visualizar o contexto abonatório de cada item lexical. A consulta confirmou a impressão, pois 19 itens são nomes de personagens: **Loredano, Ubirajara, Itaquê, Poti, Pojucã, Jurandir, Araquém, Jaguarê, Jandira, Cubi, Irapuã, Soeiro, Simões, Camacã, Canicrã, Lauriana, Jacaúna, Peri, Martim**.

Ainda seguindo a pesquisa por meio do contexto abonatório, encontramos dez itens que nomeiam as tribos e são usadas também como adjetivos para especificar um personagem**: abaré, aimoré**, **moacaras**, **pitiguaras, pitiguara, tabajaras, araguaia, araguaias, tocantim**. Outras palavras do universo dos indígenas também estão entre as 20 mais chave com zero frequência, como **tupã**, **pajé**, **guerreiros**, **cabana**, **clavina**, **relva**, **seta**. Por fim, há também as formas verbais como **tinha-se, tornou-se, sabeis, quereis, murmurou, ouviu-se, lembrou-se, dirigiu-se, ergueu-se** e os substantivos **esposo, distância** e **sois,** esta última, grafada, no texto, sem o acento.

Ressaltamos que, embora estejamos analisando as 20 primeiras palavras com frequência zero, ordenadas por chavicidade, elas não são coincidentes em todas as listas de palavras geradas, conforme se constata nas Figuras 31 a 34. Porém, a lista das palavras nos mostrou que somando a totalidade dos itens que estão entre os vinte primeiros, considerando os quatro *corpora* de referência, com frequência zero, totalizam 48. Desses 48 itens, 19 são nomes de personagens, 10 nomeiam tribos e sete referem-se ao universo dos indígenas, ou seja, 36 itens são relacionados ao contexto indígena de um total de 48 palavras.

Pode-se aventar, então, que Alencar utiliza um léxico específico em suas obras indianistas, pois a porcentagem de palavras com frequência zero nos *corpora* de referência do mesmo autor, como também o contraste com os demais *corpora*, nos comprova esta produtividade e essa especificidade do léxico utilizado por Alencar em suas obras indianistas.

Durante a análise das palavras-chave do *corpus* de estudo e das palavras de frequência zero nos *corpora* de referência em relação ao *corpus* de estudo, como leitora das obras de Alencar, percebi que muitas palavras conhecidas e recobradas pela memória não constavam das listas geradas. Então, baseado no princípio impressionístico apontado por Berber Sardinha (2004; 2009), e considerando que o *corpus* pode guiar e guia, a partir de determinado momento, a pesquisa, resolvemos fazer o levantamento das palavras genuinamente indígenas das obras de Alencar. Isso porque as listas geradas até o momento nos mostraram a

produtividade do autor em relação ao léxico específico utilizado nas obras indianistas, entretanto não extraímos o léxico especificamente indígena. Assim sendo, na seção seguinte, apresentamos o léxico indígena característico das obras indianistas de José de Alencar, que compõem o nosso *corpus* de estudo.

5.3 Vocábulos indígenas do *corpus* de estudo

Nesta seção, apresentamos os vocábulos indígenas identificados no *corpus* de estudo. Esse apanhado é importante, porque nos ajuda a levantar os dados do *corpus* que podem nos auxiliar na elaboração do Vocabulário do léxico indianista de José de Alencar. Pareceu-nos, em primeiro momento, que as palavras-chave apontavam para os nomes próprios em preponderância, porém o foco desta pesquisa não são apenas os nomes próprios, daí a necessidade de extração de todas as palavras substantivos e adjetivos indígenas. Ressaltamos que não é objetivo desta pesquisa analisar os verbos, os quais poderão ser investigados em outra oportunidade.

Ressalvamos que, ao iniciarmos a análise, por meio da observação das palavras-chave, percebemos que muitos vocábulos já conhecidos por mim, por ser leitora das obras de Alencar e já ter desenvolvido pesquisa no Mestrado sobre esse autor, não apareciam nas listas de palavras-chave. Em razão disso e no intuito de tentarmos a extração do léxico indianista, alteramos o valor de significância estatística para p= 0,00000001 e a frequência também foi alterada para o mínimo 2. Com esse ajuste nas configurações, inclusive palavras que registrassem duas ocorrências no *corpus* de estudo poderiam chegar a ser palavras-chave. Essa mudança alterou significativamente as listas de palavras-chave geradas. Consideramos essa alteração necessária e procedente, uma vez que Alencar mostra-se produtivo em termos de diversidade lexical em suas obras indianistas comparadas com as demais suas obras.

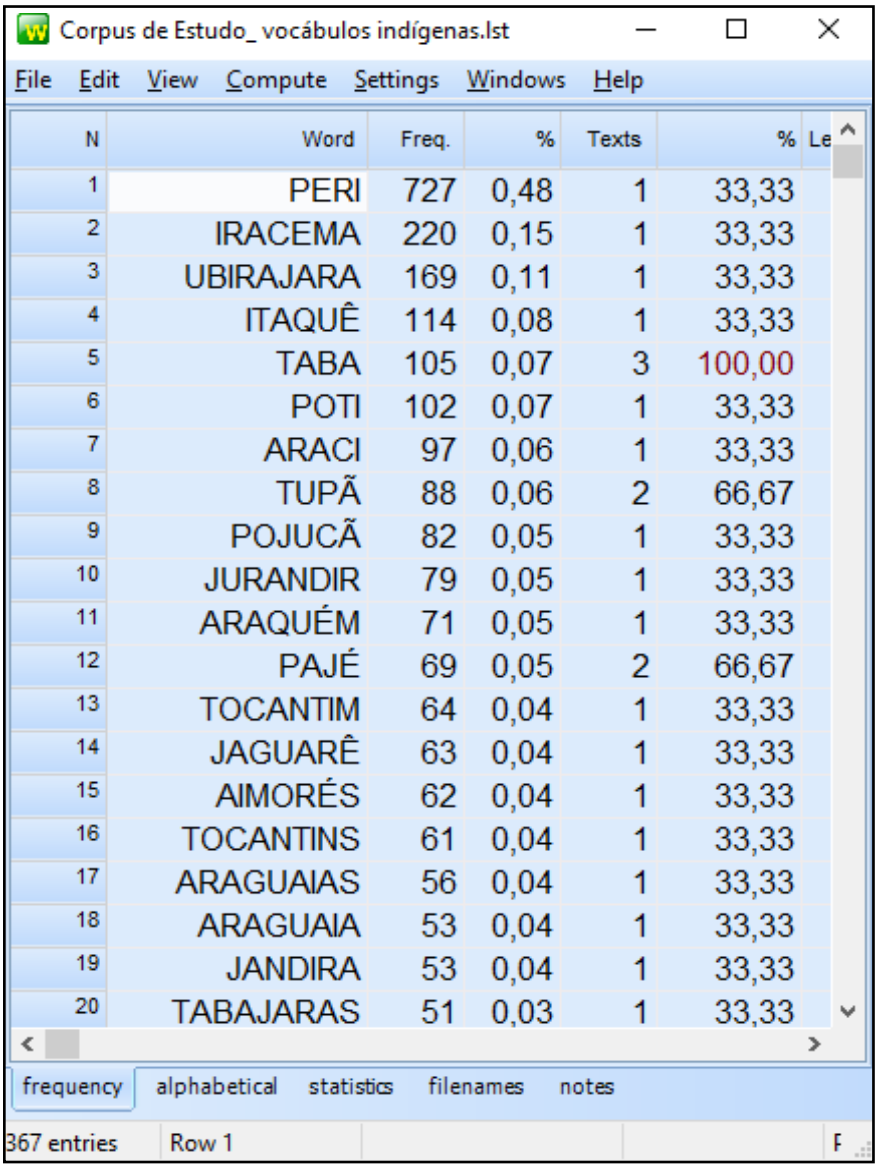
Paralelamente, utilizamos a ferramenta *WordList,* a fim de fazermos o cotejo das palavras indígenas, junto às quatro listas de palavras-chave já obtidas. A partir da *WordList,* realizamos o procedimento de limpeza do *corpus* eliminando, manualmente, todos as palavras gramaticais e formas verbais. Adotamos fazer a limpeza manualmente, porque foi importante para visualização das palavras no momento da exclusão[38], para aproximar o olhar para a lista

[38] Os procedimentos de limpeza foram, detalhadamente, descritos no Capítulo 4 "*Corpus* e Metodologia".

de palavras do *corpus* de estudo. Esse foi um procedimento que, como analistas, adotamos, fundamentalmente, em função do propósito de redigirmos os verbetes do léxico indianista para o vocabulário de consulta *online*.

A Figura 35 mostra as 20 primeiras palavras consideradas indianistas obtidas após os procedimentos de exclusão dos itens não indianistas. É possível notar que as palavras, na lista gerada, estão dispostas em ordem de frequência, da mais frequente para a menos frequente. Não é possível visualizar todos os itens nesta lista, porém a ferramenta disponibiliza as palavras de acordo com diferentes critérios: inicialmente, apresenta as palavras em ordem decrescente, cuja frequência é diferente; em seguida, a lista das palavras de mesma frequência, as quais, neste caso, são dispostas em ordem alfabética. Em síntese, a ferramenta considera como primeiro critério a frequência e, como segundo critério, a ordem alfabética.

Figura 35: Vista parcial das 20 primeiras palavras indianistas do *corpus* de estudo



N	Word	Freq.	%	Texts	%	Le
1	PERI	727	0,48	1	33,33	
2	IRACEMA	220	0,15	1	33,33	
3	UBIRAJARA	169	0,11	1	33,33	
4	ITAQUÊ	114	0,08	1	33,33	
5	TABA	105	0,07	3	100,00	
6	POTI	102	0,07	1	33,33	
7	ARACI	97	0,06	1	33,33	
8	TUPÃ	88	0,06	2	66,67	
9	POJUCÃ	82	0,05	1	33,33	
10	JURANDIR	79	0,05	1	33,33	
11	ARAQUÉM	71	0,05	1	33,33	
12	PAJÉ	69	0,05	2	66,67	
13	TOCANTIM	64	0,04	1	33,33	
14	JAGUARÊ	63	0,04	1	33,33	
15	AIMORÉS	62	0,04	1	33,33	
16	TOCANTINS	61	0,04	1	33,33	
17	ARAGUAIAS	56	0,04	1	33,33	
18	ARAGUAIA	53	0,04	1	33,33	
19	JANDIRA	53	0,04	1	33,33	
20	TABAJARAS	51	0,03	1	33,33	

Fonte: A autora, por meio da ferramenta *WordList* do *WST*

Por meio do contraste entre a lista de palavras do *corpus* de estudo e as quatro listas de palavras-chave, foi possível observar que, apesar de o ajuste ter sido feito para frequência 2, alguns vocábulos indianistas de frequência 1 não haviam sido capturados pela ferramenta *KeyWords*. Em razão disso, entendemos que, para os objetivos desta pesquisa, a extração dos vocábulos indianistas deveria ser complementada a partir desse estudo paralelo ainda mais fino, em que pela ordem alfabética da lista de palavras foram identificados os vocábulos indianistas ausentes nas listas de palavras-chave.

Ao encerrarmos essa nova etapa de apuração do léxico, a lista de vocábulos indianistas aumentou em mais 165 palavras utilizadas apenas uma única vez pelo autor, ou seja, que registraram apenas uma ocorrência no *corpus* de estudo. Após a consulta a todos os dicionários de exclusão, já somados os 165 vocábulos de ocorrência 1 (hápax legomena), alcançamos 367 considerados indianistas. Ressaltamos que estes 165 itens também comporão o vocabulário *online* com os termos indianistas[39].

O Quadro 9 apresenta a lista das 367 palavras indígenas identificadas no *Corpus* de estudos, com base nos dicionários consultados, acompanhadas da frequência.

Quadro 9: Lista das 367 palavras consideradas indígenas do *corpus* de estudo

Palavra	Frequência	Palavra	Frequência	Palavra	Frequência	Palavra	Frequência	Palavra	Frequência
abacaxis	1	abaeté	2	abaré	1	abati	2	açaí	2
acajás	1	acaraú	7	acaris	1	acarú	1	acauã	1
aguapés	1	aimoré	7	aimorés	62	aimóres	1	aipim	1
airi	1	anajê	3	ananás	6	andira	10	andiroba	1
angico	2	anhangá	3	anum	1	apodi	1	ará	4
araçá	1	araçás	1	aracati	1	araci	97	araçóia	2
araguaia	53	araguaias	56	araquém	71	arara	8	ararê	22
araribás	1	araribóia	2	araruna	1	aratanha	1	aratuba	1
aresqui	1	arirama	1	ariranha	1	aroeira	1	aroeiras	1
arroio	2	bambu	1	ati	1	atiati	2	balaio	1
banana	1	bananas	1	bananeiras	2	batuireté	5	biaribi	2
biribá	1	boicininga	2	boitatá	1	boré	2	borés	3
cabuíba	1	cacique	20	caiçara	4	caiçaras	1	caiporas	1
caitetus	1	caititus	1	cutia	5	cajazeira	3	caju	6
cajueiro	4	cajueiros	1	cajus	1	camacã	29	camoropim	2
camorupim	1	campeiro	2	campinas	2	camuci	1	Camucim	15
camucins	2	canicrã	18	caninana	1	capim	2	capivara	3
capoeira	1	caraíba	1	caramuru	1	caramurus	1	caraúba	1
carcará	1	carimã	1	carioba	2	carnaúba	7	catuíba	1
cauã	4	cauatá	1	caubi	46	cauim	12	ceará	3
graúna	2	cearense	2	ceci	31	choça	2	cipó	7
cipós	5	clavina	22	coati	1	coatiabo	6	copaíba	1
jetaí	2	corisco	1	coruja	4	crajuá	1	crajuru	2

[39] Ressaltamos que estes 165 itens também comporão o vocabulário *online* com os termos indianistas. Sendo assim, resolvemos mantê-los na lista de vocábulos indianistas do autor.

crauatá	1	craúba	3	crautá	3	craviri	1	crebã	2
croás	1	cuandu	1	cumari	2	cumbucas	1	cupim	1
curumim	7	embaíba	3	emboabas	2	gará	1	goaná	1
goiamum	3	goitacá	9	goitacás	7	guabiroba	1	guaiúba	1
guanumbi	5	guará	3	guaraciabas	4	guaraná	1	guarani	4
guaranis	3	guaribu	5	guaximas	3	guaxinim	1	iara	2
ibiapaba	4	ibiapina	1	icó	1	igaçaba	3	igaçabas	2
igapê	1	igara	6	igaras	3	iguape	1	imbé	1
imbu	2	inhuma	4	intanha	1	inúbia	17	inúbias	1
ipu	7	ipus	1	iracema	220	irapuã	44	irara	4
irerê	1	itaoca	2	itaquê	114	itaquêe	1	jaburu	1
jabuti	2	jacamim	14	jaçan	1	jaçanã	6	jaçanãs	2
jacarandá	9	jacarandás	1	Aquiraz	1	jacarecanga	3	jacareí	1
jacaúna	28	jaci	1	jacobina	1	jacus	1	jaguar	15
jaguaraçu	2	jaguarê	63	jaguaribe	3	jandaias	1	jandira	53
japi	4	japim	1	jararaca	1	jataí	1	jati	2
jatobá	13	jatobás	1	javari	4	jenipapo	3	jequiriti	1
jequis	1	jereraú	1	jetica	1	jibóia	6	jibóias	1
jirau	5	juazeiros	3	juçara	6	juquiri	1	jurandir	79
jurema	79	juriti	4	jurupari	1	juruti	2	jutaí	1
jutorib	1	macana	1	mairi	3	majé	17	majoí	1
manacá	4	manam	1	manava	2	mandioca	2	mangaba	1
mangabeira	1	maniva	3	marabá	1	maracá	7	maracajá	1
maracás	1	maracatim	4	maracujá	5	maranduba	5	maranguab	6
mearim	13	membi	1	meruoca	1	mingau	1	moacaras	21
moacir	2	mocejana	3	mocoribe	9	moquém	1	morubixaba	5
muçurana	3	mundaú	2	murinhém	7	muriti	1	muritiapuá	1
muritis	1	mururê	1	mutucas	1	nambu	6	nandu	2
nhengaçara	1	nhengaçaras	3	nhengaíba	1	oca	7	ocara	8
ogib	1	oitibó	3	oticica	1	ouricuri	1	paca	1
pacas	1	pacatuba	1	pacoti	1	pahã	7	pajé	69
pajés	3	papagaios	2	paquequer	20	pará	2	paraíba	14
quatis	2	parnaíba	1	pequiá	1	peri	727	piaçaba	1
piau	1	piracém	1	pirajá	6	piranhas	1	pirapora	1
pirijá	1	pirogas	2	pitanga	2	pitiguara	33	pitiguaras	37
pocema	9	pojucã	82	porongaba	6	poraquê	1	pororoca	1
potengi	1	poti	102	potiguara	3	potiguaras	2	quixeramobim	1
sabiá	6	sabiás	1	sagui	1	saí	2	saixê	1
sapé	3	sapiranga	1	sapopema	2	sapoti	1	sapucaia	3
saúva	4	saúvas	1	sopé	1	sucuri	4	taba	105
tabajara	41	tabajaras	51	tabas	13	taboca	1	tacape	38
tacapes	9	taioba	1	tamandaré	4	tamanduá	2	tamoios	1
tanatinga	1	tangapema	6	tapir	13	tapuia	17	tapuias	15
tapuitinga	2	taquara	1	taquari	1	tatu	1	tauape	1
teiú	1	ticum	1	tocantim	64	tocantins	61	traíra	1
traíras	1	trairi	3	trocano	6	tubim	3	tucano	10
tucanos	1	tuins	1	tupã	88	tupi	2	tupinambá	3
tupinambás	9	ubaia	3	ubirajara	1	ubirajaras	1	ubiratã	8

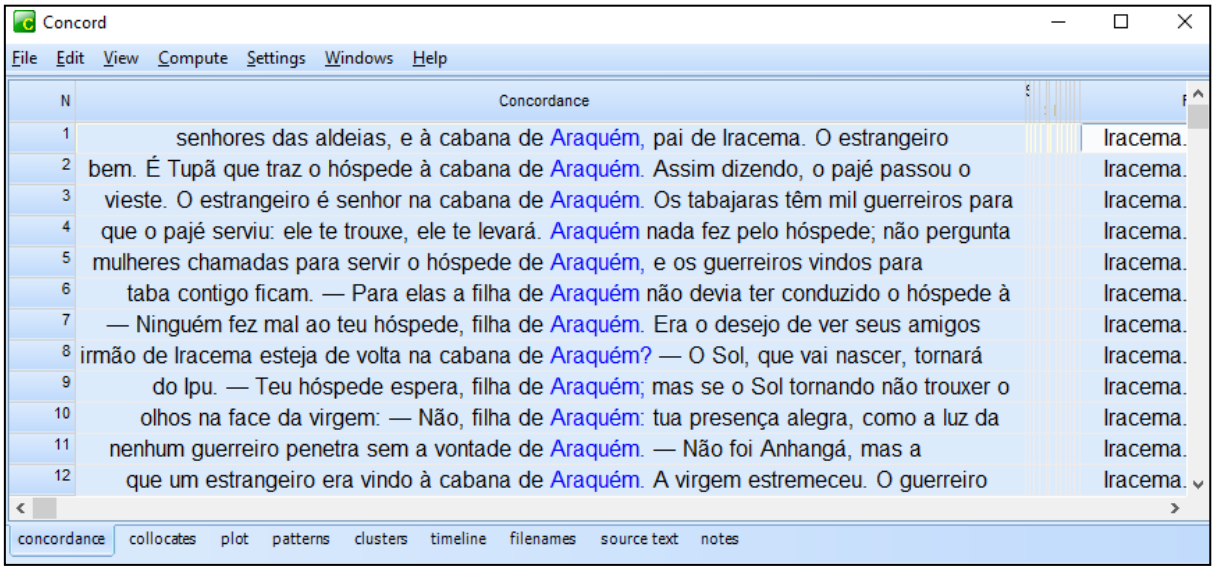
uiraçaba	3	uiraçu	5	uraçá	1	urataí	1	uru	4
uruburetama	1	urubus	2	urus	1	urutau	2	uvaias	1
xingu	1	zabelê	2						

Fonte: A autora

Consideramos um número absoluto de ocorrências de 367 palavras indígenas, para o número de 758 substantivos e adjetivos presente em todo o *corpus* de estudo, muito relevante e capaz de confirmar a premissa inicial de que Alencar utiliza um léxico específico em suas obras indianistas. Em porcentagem, o valor aproximado é de 41%, ou seja, perto da metade dos substantivos e adjetivos são indígenas.

Um outro fator relevante, observado durante a pesquisa, é o número de palavras que não constam em nenhum dos dicionários consultados. Um total de 57 palavras são utilizadas pelo autor e não figuram em nenhum dos dicionários de exclusão. Para se chegar a este número, consultamos as 758 palavras candidatas a vocábulos indígenas em todos os dicionários de exclusão: GD, TS, AGC, LCT, LMSP, AMS e RB e, desse número, obtivemos as 57 que não constam em nenhum deles. Consideramos palavras indígenas a partir da análise do contexto por meio da ferramenta *Concord* do *WST*. A Figura 36, exemplifica uma das palavras, **Araquém**, utilizadas pelo autor que foram consideradas indígenas, embora não tenha sido identificada em nenhum dos dicionários de exclusão. A palavra **Araquém** é o nome do pai de **Iracema**, portanto, trata-se de um personagem indígena.

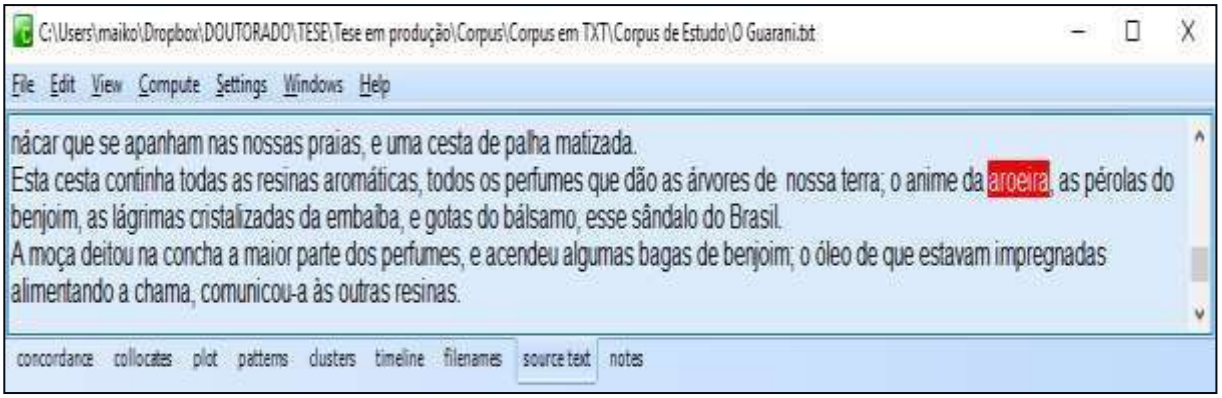
Figura 36: Linhas de concordância da palavra **Araquém**



Fonte: A autora, com base na ferramenta C*oncord* do *WST*

Como investigamos cada uma das palavras, selecionamos a primeira da lista por ordem alfabética, que é **Araquém**, para demonstração de que se trata de uma palavra utilizada pelo autor com significado de uma indígena. Como é um nome de personagem indígena, assim como **Ararê** e **Aresqui**, apresentamos também um outro trecho em que outro vocábulo, **aroeira**, que não seja antropônimo, seja citada por Alencar, com o sentido indígena, conforme Figura 37. Este trecho foi obtido também por meio da ferramenta *Concord* do *WST*. Notamos que ele se refere a um elemento da flora e é utilizado como elemento na composição da descrição realizada por Alencar, ao lado de outros elementos também da flora como **árvores**, **resinas**, **benjoim**, **embaúba** e **bálsamo**. Ressaltamos que consideramos esta palavra indianista, porque ela não consta dos dicionários consultados de língua geral. As palavras elencadas também não estão registradas nos dicionários de língua indígena, porém apresentam um sentido que corrobora o contexto indígena das obras.

Figura 37: Linhas de concordância da palavra **aroeira**



Fonte: A autora, com base na ferramenta C*oncord* do *WST*

Como dito, esse procedimento foi repetido com todas as palavras e, com isso, obtivemos as 57 palavras não dicionarizadas, porém consideradas indianistas por fazerem parte do universo dos indígenas de Alencar. No Quadro 10, constam as 57 palavras, acompanhadas pela frequência em que aparecem no *corpus* de estudo.

Quadro 10: Palavras consideradas indígenas que não constam dos dicionários de exclusão

Palavra não constante dos dicionários	Frequência	Palavra não constante dos dicionários	Frequência	Palavra não constante dos dicionários	Frequência	Palavra não constante dos dicionários	Frequência	Palavra não constante dos dicionários	Frequência
Araquém	71	Ararê	22	Aresqui	1	Aroeira	2	Arroio	2
Balaio	1	Banana	1	Bananas	1	Campeiro	2	Canicrã	18
Carcará	1	Cauatá	1	Caubi	46	Choça	2	Clavina	22
Coatiabo	6	Corisco	1	Coruja	4	Crajuá	1	Crajuru	2
Crautá	3	Craviri	1	Crebã	2	Croás	1	Gará	1
Goaná	1	Goiamum	3	Guabiroba	1	Guaiúba	1	Guaraná	1
Itaquê	114	Jutorib	1	Macana	1	Manam	1	Manava	1
Murinhém	7	Muritiapuá	1	Nhengaçara	4	Nhengaiba	1	Ogib	1
Oitibó	3	Pahã	7	Papagaios	2	Porangaba	6	Potengi	1
Sapoti	1	Sopé	1	Tanatinga	1	Tapuitinga	2	Tauape	1
Ticum	1	Tocantim	125	Trairi	3	Ubirajara	169	Ubirajaras	1
Ubiratã	8	Uraçá	1						

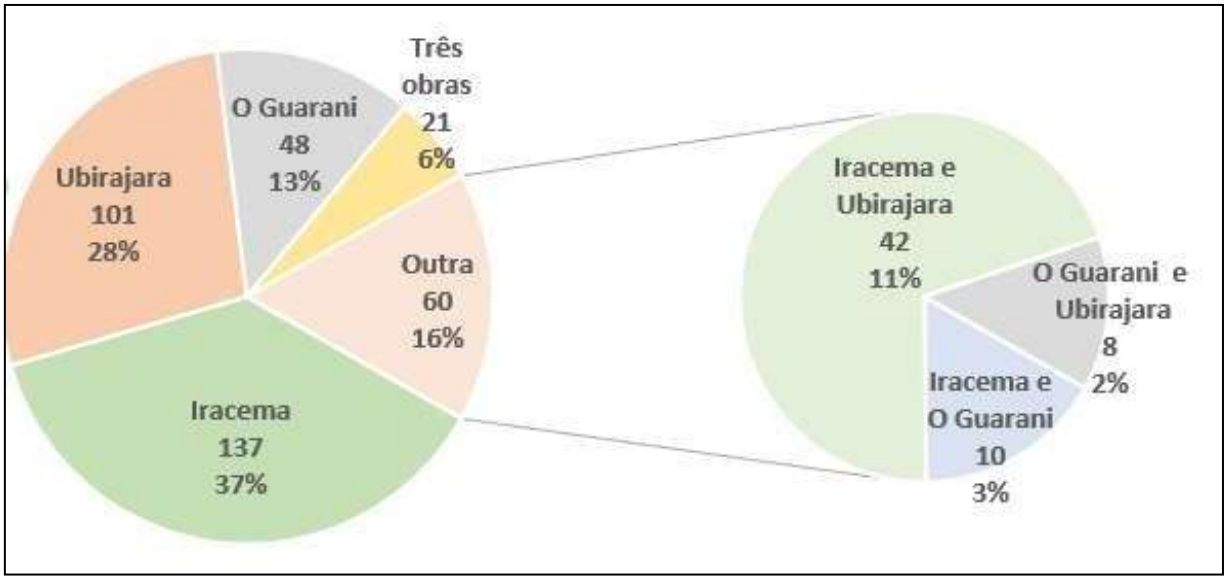
Fonte: A autora

Ainda analisando os dados encontrados, verificamos que, das 57 palavras que não constam dos dicionários, 29 são empregadas uma única vez pelo autor em suas obras indianistas[40]. Esse quadro ratifica a comprovação de que Alencar diversificou, sobremaneira, o léxico em suas obras indianistas.

Além dos dados apresentados, empreitamos uma nova pesquisa, a fim de investigar também em qual obra Alencar teria a veia criativa mais aguçada. Então, por meio da ferramenta *Concord* do *WST*, pesquisamos as 367 palavras indianistas para verificar em que obra, ou em quais obras, Alencar utilizou cada uma delas. O Gráfico 1 apresenta os dados organizados em que os números indicam a quantidade de palavras utilizadas, exclusivamente, em cada obra ou em cada par de obras ou a utilização dos vocábulos indianistas nas três obras. A porcentagem representa a distribuição, portanto, das 367 palavras indianistas identificadas no *corpus* de estudo.

[40] Mantivemos a grafia das palavras no singular e no plural, em razão de, em alguns casos, o plural referir-se a elemento distinto do singular.

Gráfico 1: Identificação da quantidade de palavras indianistas nas obras de Alencar



Fonte: A autora

É possível notar que a maior criatividade lexical de Alencar, considerando o léxico indianista, está em *Iracema*, pois dos 367 vocábulos indígenas, 137 são utilizadas apenas neste romance, ou seja, 37% do total dos vocábulos. Seguindo a mesma análise, Alencar utiliza 101 vocábulos apenas em *Ubirajara*, o que corresponde a 28%; e em *O Guarani*, o autor utiliza 48 dos 367 vocábulos, correspondendo a 13%. Considerando os vocábulos que Alencar utiliza nos três romances, há 21 vocábulos, o que totaliza seis por cento do total.

O gráfico desmembra no círculo menor, as palavras utilizadas em duas das obras do autor. O maior índice de repetição de palavras está entre os romances *Iracema* e *Ubirajara*, pois totaliza 42 vocábulos, isto é, 11% do total das 367. Alencar utiliza tanto em *Iracema* quanto em *Ubirajara.* Considerando os romances *Iracema* e *O Guarani*, Alencar utiliza 10 vocábulos, correspondente a três por cento do total. Por fim, o autor utiliza 8 vocábulos em *O Guarani* e *Ubirajara*, correspondendo a dois por cento também do total.

Analisando o número de vocábulos indianistas que Alencar utiliza em cada obra, verificamos que, se considerarmos os vocábulos utilizados apenas em *Iracema* que são 137; somando-se aos utilizados nas três obras, que são 21; mais os utilizados em *Iracema* e *Ubirajara* que são 11 vocábulos; acrescidos dos 10 vocábulos utilizados em *Iracema* e em O *Guarani*, totalizamos 210 vocábulos indianistas são utilizadas em *Iracema*, das 367 totais, um total de 57%.

Em seguida, podemos dizer que a criatividade do autor se manifesta em *Ubirajara*, pois emprega 101 vocábulos apenas neste romance. Considerando também os 21 que se repetem nas três obras; mais os que são empregados ao mesmo tempo em *Ubirajara* e *Iracema,* que são 42 vocábulos, seguidos de 8 que emprega em *Ubirajara* e *O Guarani*, totaliza, então, 172 vocábulos do total de 367 empregados no romance *Ubirajara*.

Finalmente em O Guarani, dos 367 vocábulos, 48 são empregados apenas neste romance. Se somarmos os que se repetem nas três obras, que são 21; mais os que são utilizados em O *Guarani* e *Ubirajara*, 8 vocábulos; e aqueles utilizados em *O Guarani* e *Iracema*, 10 vocábulos, totalizamos 87 vocábulos utilizados no romance *O Guarani* das 367 que correspondem ao total de indígenas.

Esses números demonstram que a produtividade lexical é prevalente no romance *Iracema*. Ressaltamos que essa prevalência não tem relação com o ano ou a ordem de publicação dos romances, pois *Iracema* foi publicado em 1865, depois de *O Guarani*, 1857 e antes de *Ubirajara*, 1874. Poder-se-ia supor, em princípio, que haveria o predomínio por palavras indígenas no romance *Ubirajara*, pois Alencar retrata a vida dos índios, nas batalhas e romances entre tribos distintas, porém isso não se confirmou. A criatividade lexical está em *Iracema* que, apesar de o palco do romance ser as terras indígenas, retrata o amor entre a índia **Iracema** e o português **Martim**, ou seja, há a presença de personagens não indígena. Por outro lado, *O Guarani* retrata um outro lado da relação branco e índio, isto é, apresenta um índio inserto no cenário constituído por branco.

Revisitando o objetivo da tese proposto neste capítulo, podemos afirmar que Alencar utiliza um léxico específico em sua tríade indianista. Esta afirmação se baseia no contraste entre o *corpus* de estudo, composto pelos romances indianistas com o *corpus* de referência também constituído pelas das obras não indianistas de Alencar o CorpRef-Alencar; assim como, por meio do contraste com os demais *corpora* de referência: CorpRef-Lácio-Web, CorpRef-AcadTeses e CorpRef-Nov.

Com base nos dados apresentados, podemos afirmar que Alencar é produtivo por utilizar um léxico específico em suas obras indianistas, tendo em vista que as palavras-chave de todas as *KeyWords* de contraste com os três *corpora* são praticamente as mesmas entre as primeiras 20 das listas. Corroborando, ao analisar as palavras com frequência zero nos *corpora* de referência também demonstrou uma quantidade de palavras significativa. Outro

aspecto foi a porcentagem de palavras consideradas indígenas em relação à totalidade de substantivos e adjetivos presentes no *corpus* de estudo. Mais de 48% de substantivos e adjetivos que o autor utiliza em suas obras indianistas, são indígenas, conferindo o caráter de produtividade e ineditismo lexical em suas obras.

Encerramos este capítulo em que analisamos o léxico das obras indianistas de Alencar em contraste com os *corpora* de referência. No capítulo seguinte, tratamos da Etimologia dos vocábulos indígenas que José de Alencar empregou em seus romances indianistas.

CAPÍTULO 6 - A ETIMOLOGIA FICCIONAL CONTEXTUAL EM ALENCAR: A CRIAÇÃO DE VOCÁBULOS INDÍGENAS

Alencar afirma que, desde cedo, quando começaram os desejos de escrever, "a raça selvagem indígena" despertava-lhe interesse. Apesar de não possuir conhecimentos suficientes para apreciar esta raça, ele se empenhou em estudos por meio de textos já publicados sobre o tema, porém não encontrou, dentre eles, uma "poesia nacional", tal como ele percebia "os selvagens". Em razão disso, Alencar procurou apropriar-se do linguajar indígena para tematizá-los em suas obras literárias.

Considerando as obras literárias de Alencar, este capítulo tem por intuito apresentar uma análise do léxico indianista desse autor sob a perspectiva da Etimologia Ficcional Contextual. Expandimos o conceito de Etimologia para Etimologia Ficcional Contextual, que consideramos a análise ou a busca da origem dos vocábulos, a partir da interpretação no contexto de emprego nas obras. Entretanto, em razão da quantidade de vocábulos indígenas identificados (367), e por razão de extensão desta tese, não apresentamos a análise de todos os vocábulos. Vale ressaltar que não será relevante analisar se Alencar utilizou uma língua *Tupi* ou *Guarani*, pois o autor faz uso de vocábulos dessas duas línguas e de outros dialetos indígenas. Interessa-nos analisar o processo criativo de Alencar ao criar e utilizar o léxico indianista.

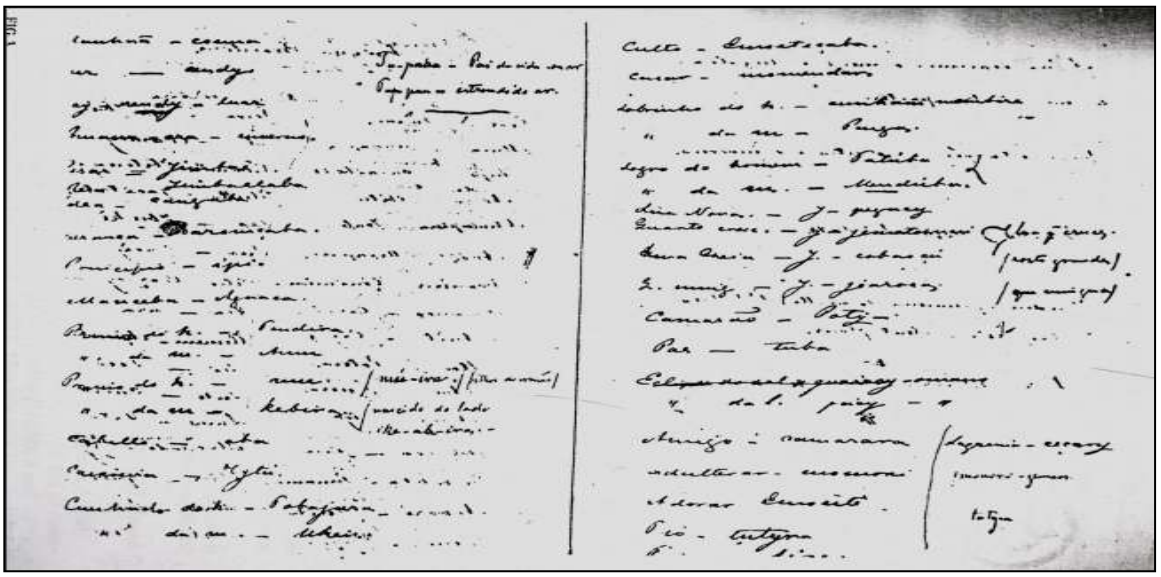
José de Alencar exibe, nitidamente, um acervo lexical indianista inegável, aspecto comprovado pelos 367 vocábulos indígenas identificados no capítulo anterior, utilizados em sua tríade: *Iracema*, *O Guarani* e *Ubirajara*. É possível afirmar também que Alencar adquiriu um vasto conhecimento sobre a língua indígena, o que se comprova ao iniciarmos uma análise pelas notas que o autor coloca no final dos romances, além das minúcias sobre a língua indígena que o autor escreve na Carta ao Dr. Jaguaribe.

Com sobriedade, Alencar escreveu sua famosa lenda indígena, cônscio das armadilhas e incoerências a que estaria sujeito, como qualquer romancista que se arriscasse a tematizar a raça indígena, no Brasil, naquele momento. Porém, ele não se acovardou e se preparou para as farpas que viriam. Essa preparação se deu em forma de estudos sobre a língua indígena que ele empreendeu.

Segundo Schwamborn (1998, p. 403), os conhecimentos de Alencar sobre a língua indígena não foram adquiridos "através de viagens ao Amazonas e demoradas permanências junto aos índios, mas frequentando bibliotecas. Alencar não foi nem general, nem padre, nem etnólogo; foi escritor de ficção". A autora afirma também que, de acordo com os livros de registro de visitação do público, Alencar passava várias horas do dia na Biblioteca Nacional do Rio de Janeiro pesquisando sobre a História do Brasil em vários livros como os de Aires de Casal e Varnhagen. Autores que teriam colaborado para que Alencar criasse o léxico indianista empregado em sua trilogia.

Além das horas que Alencar passava estudando na Biblioteca Nacional, o que alargou seus conhecimentos sobre a língua e costumes indígenas, Schwamborn (1998, p. 415) afirma que Alencar também fazia anotações sobre a língua indígena que estão conservadas no Museu Histórico[41]. No Caderno X de seus *Apontamento Diversos*, Alencar escreveu uma lista de vocábulos, incluindo prefixos, sufixos e radicais, que traz como título *Lingua Basilica*. A autora teve contato com esse caderno e afirma que "a maioria dos vocábulos são tirados daí e foram utilizados por Alencar em *Iracema*"[42]. Embora ilegível, em razão do processo de envelhecimento natural das folhas, na Imagem 3, é possível visualizar uma folha do Caderno X do José de Alencar.[43]

Imagem 3: Folha do Caderno X dos *Apontamentos Diversos* de José de Alencar



[41] Schwamborn (1998), em nota na página 418, afirma "O saudoso pesquisador e guardião da memória de Alencar, Prof Fabio Freixieiro que me mostrou estes cadernos e também esta lista em 1983".
[42] Houve um contato com o Museu Nacional para resgatar esse material, porém, como se trata de um acervo de grande relevância para a História da Língua Portuguesa, o acesso é rigorosamente restrito.
43 Estes cadernos estão no Museu Histórico Nacional do Rio de Janeiro.

Fonte: Schwamborn (1998, p. 593)

Reproduzimos também uma lista de alguns exemplos apresentados por Schwamborn (1998, p. 415-416), transcritos do mencionado Caderno de Alencar:

Deus –	Tupan	Terra -	Iby
Alma -	anga	Sol -	Coaracy
anjo -	caribéte	Nuvem -	Ibitinga
Ar -	ibitu	Estrela -	jacitata
Atmosfera -	Ibaté	Ceo -	Ibaké
Arco -	Uirapara	Lua -	Jacy
Arco-iris -	Guaiamim	Estio -	Coaracy-ara
Branco -	Cariba	Dia -	Ara
Anil -	Cayby	[ilegível] -	Apiaba
Flor -	Potira	Lei -	Tecô
Ave -	Guira	Leito -	Cambi
Agua -	Ig	Lume -	Tata
Senhora -	Iara	Machado -	Gy
Borboleta -	Panamá	Mez -	Jacy
Arvore -	[ilegível] iba	Mar -	parana
Pena -	Cacy	Mel -	ira

Além dessa lista, a autora reporta a outras páginas no Caderno de Alencar e cita as seguintes anotações do autor:

Cunhado do h. – Tobajara

Camarão – Poty

Tu – paya (Tupum) pae da vida

Tu – ipys (Tupis) primeiros q viverão

Guaranys – Igara –ne-y – que hao de ser os senhores das aguas

Ibyra – Iby-ira – folhas da terra – arvores

Iguara – ig – uaras – senhora d'agua – canoas

Ta – puya – taba – fu[n]gir – que abandonou as Tabas – bábaro – selvagens

(SCHWAMBORN, 1998, p. 416)

Outra anotação relevante no Caderno X de *Apontamento Diversos* mencionada por Schwamborn (1998), é uma lista em que Alencar apresenta uma comparação entre algumas palavras em três línguas: *Guarany*, *Tupi* e *Omagua*. Fato que pode comprovar que o autor estudou outras línguas indígenas, além do *Tupi* e oscila entre explicações de vocábulos oriundos do *Tupi* e do *Guarani*.

Iniciamos nossa análise com o nome da personagem principal que também nomeia o romance: **Iracema**. Há algumas tentativas de explicar a etimologia do vocábulo **Iracema**, dentre elas, o próprio Alencar, em nota, ao final do romance, afirma que "**Iracema** - Em guarani significa lábios de mel - de *ira*, mel e *tembe* - lábios. *Tembe* na composição altera-se em *ceme*, como na palavra *ceme-iba*". Já Navarro (2013) não menciona a língua guarani e propõe que **Iracema**, vem do nhengatu (uma língua amazônica), e foi utilizado por Alencar com o significado original (*ira*, abelhas + *sema*, saída) e alterada para "lábios de mel". O dicionário de GD não possui o verbete **Iracema** como antropônimo, mas traz as palavras utilizadas por Alencar no processo de composição *ira* ou *yra* – mel (p.505); e *tembê* – beiço (ALENCAR, 1965, p. 489).

A fim de verificar a datação da primeira ocorrência do vocábulo **Iracema** em textos escritos no português brasileiro, verificamos no *Corpus do Português* (DAVIES, 2016). Apesar de os dados apresentados neste *corpus* não serem a garantia de que o vocábulo de fato ocorreu, pela primeira vez, no texto compilado para formação do *Corpus do Português*. Isso porque poderia haver alguns textos, cuja existência é desconhecida, porém nos basearemos nele para nossas análises, pois trata de um *corpus* robusto em termos de quantidade textos compilados e em número de palavras registradas.

Vale ressaltar que a ordem de aparição dos textos no Corpus do Português (DAVIES, 2016), versão histórica, não atende ao critério de cronologia. Para verificar a datação dos textos utilizamos as linhas de concordância em que aparecem o ano do texto em que o vocábulo foi utilizado. A Figura 38 nos apresenta uma vista parcial do vocábulo **Iracema** do *Corpus* do Português.

Figura 38: Linhas de concordância do vocábulo **Iracema**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

É possível notar que a primeira linha com o registro do vocábulo **Iracema** é de um texto de Augusto do Anjos, porém esse autor é um sucessor de José de Alencar. A Figura 39 mostra os detalhes sobre a obra *Eu* de Augusto dos Anjos, que foi publicada em 1912, portanto posterior a *Iracema* de Alencar de 1865. A explicitação das datas de publicação dos textos dos autores comprova que o texto de Augusto dos Anjos, mesmo não constando esse registro no *Corpus do Português*, é posterior e faz referência à índia personagem de Alencar.

Figura 39: Vocábulo **Iracema** em contexto ampliado no texto *Eu* de Augusto dos Anjos

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

A Figura 40 apresenta detalhes de uma edição de *O Guarani* e que, provavelmente, não corresponderia à primeira edição, uma vez que faz referência ao romance *Iracema*.

Figura 40: Vocábulo **Iracema** em contexto ampliado no livro *O Guarani* de José de Alencar

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Cabe destacar, também, que nos dicionários RB, AMS e LMSP não há registro do vocábulo **Iracema**, já GD traz o verbete **Iracema** referindo-se a uma cidade do Ceará. Outros aspectos importantes em relação ao vocábulo **Iracema** é que: primeiro, Navarro (2013) discorda de Alencar em relação à origem do nome, afirmando ser da língua nhengatu, já que Alencar afirma ser de origem guarani. Segundo, Navarro afirma que Iracema tem o significado original *ira* – abelha e *sema* – saída, ou seja, seria a saída das abelhas, ou seja, o próprio mel, porém Alencar deixou clara a etimologia que pretendia para o nome da heroína indígena era *ira* – mel e *tembe* – lábios: a virgem dos lábios de mel. É possível afirmar que Alencar compôs o vocábulo **Iracema** a partir da definição de GD, porém atribuiu o significado pretendido por ele.

Assim sendo, com base nos documentos consultados, o vocábulo **Iracema** é utilizado pela primeira vez por Alencar, como antropônimo. **Iracema** é, portanto, um étimo criado por José de Alencar, trata-se, portanto, de uma etimologia ficcional contextual.

Outra linha de análise para o vocábulo **Iracema** foi proposta por Afrânio Peixoto, em 1931, que descobriu nas letras do nome **Iracema** o anagrama de **América**. Sobre este aspecto, há aqueles que dizem que **Iracema** representa a criação da **América** virgem, porém não nos ateremos a esta questão, tendo em vista que o próprio autor não estabelece relação entre os vocábulos **Iracema** e **América**.

Alencar coloca **Iracema** diante da imaginação do leitor por meio das descrições em forma de comparação "Iracema, a virgem dos lábios de mel, que tinha os cabelos mais negros que a asa da graúna e mais longos que seu talhe de palmeira". E é essa a imagem que Alencar deseja para sua heroína, a companhia doce e servil ao seu amor, Martim; chama-a de "virgem", uma condição importante para a mulher indígena na visão das tribos. A virgindade da índia é "doada" também para o jovem português em demonstração de amor incondicional. Passa a ser considerada uma espécie de santa nacional e referência no Brasil ao se utilizarem, em contextos diversos, expressões como "terra de Iracema" e "filhos de Iracema", as quais, inicialmente, designavam os que nasciam no Ceará, porém, tempo depois, denominou a todos os brasileiros.

Iracema, a virgem dos lábios de mel, conquistou a todos e exerceu nos seus tempos futuros um efeito não pensado por Alencar. Seu nome já foi utilizado extensivamente para nomear as meninas que nasciam no Ceará e, em menor proporção, nos outros cantos do

Brasil. Nomeia também praias, estabelecimentos comerciais de todos os tipos, ruas, enfim saiu do livro para ganhar notoriedade no país.

Tamanha é a representatividade de **Iracema** que, em 2011, foi estabelecido o dia de **Iracema**, 1º de maio. O marco foi criado pela Lei Nº 9.884/2011[44], para prestigiar a personagem e como uma forma de fomentar ações que difundam o romance e colaborem com a manutenção da memória da índia que representa, literariamente, a formação do estado do Ceará.

O direcionamento de Alencar para compor seus personagens e, imbricado de ambições ideológicas e literárias, cria também o personagem **Moacir**. Não é um personagem determinante no romance, porém não é de miudeza o seu valor. Pode-se dizer que o enredo é que direciona o autor a criar e a nomear o personagem **Moacir**, filho de **Iracema** e **Martim**.

Alencar traz, na seção de notas do livro *Iracema,* a descrição "**Moacir** — Filho do sofrimento: de *moaci* — dor, e *ira* — desinência que significa — saído de". Devido à necessidade de explicitar o emprego dos vocábulos em seus romances, o autor cria o étimo ao propor um processo de formação e um significado ao vocábulo para designar, por meio do nome, o que desejava para o filho de **Iracema**, ou seja, aquele que nasceu em decorrência do sofrimento físico e emocional da mãe. Além de ressaltar no próprio romance pela fala de **Iracema**, "- Tu és Moacir, o nascido do meu sofrimento" o autor, para reforçar a intenção de explicar que **Moacir** é fruto de um amor que trouxe sofrimentos à mãe, também traz a nota de composição do vocábulo para garantir seu propósito.

A fim de corroborar a análise proposta, buscamos o vocábulo **Moacir** no C*orpus do Português* em sua versão histórica, que confirma a primeira ocorrência no romance **Iracema**[45], conforme Figura 41.

[44]Dados publicados pelo jornal Diário do Nordeste. Disponível em: <http://diariodonordeste.verdesmares.com.br/cadernos/cidade/dia-de-iracema-memoria-da-india-permanece-viva-1.1931681>. Acesso em: 05 maio 2018.

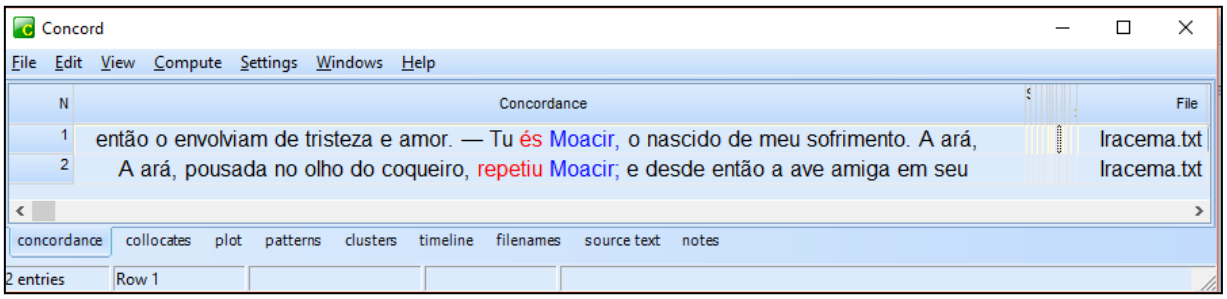
[45] Para esta afirmação, consultamos as linhas de concordância, a fim de verificar a datação dos textos.

Figura 41: Linhas de concordância do vocábulo **Moacir**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Alencar, portanto, desejava que o nome do filho de **Iracema** fosse um reflexo de sua vida durante a gestação e o parto: sofrimento. A Figura 42, nos mostra as linhas de concordância da palavra **Moacir** no *corpus* de estudo, corroborando a explicitação de Alencar sobre o significado do nome do filho de **Iracema**.

Figura 42: Linhas de concordância do vocábulo **Moacir**



Fonte: A autora, a partir da ferramenta *Concord* do *WST*

A obra de Alencar teve uma repercussão tão expressiva que a palavra **Moacir**, apesar de apenas duas ocorrências no romance, foi incorporada à Língua Portuguesa e ganhou tamanha popularidade que, a partir da publicação de *Iracema*, foi utilizada para batizar meninos. É possível que muitos pais que utilizam Moacir para dar nome a seus filhos, desconheçam o significado pretendido por Alencar no romance. Assim sendo, a popularidade do personagem preponderou sobre o significado.

Ao analisar as explicações de Alencar, em suas notas, por exemplo da palavra **Iracema** que, pela composição proposta "Em guarani significa lábios de mel – de *ira*, mel e *tembe* – lábios. *Tembe* na composição altera-se em *ceme*, como na palavra *ceme iba*.", ou seja, **Iracema** significa *lábios de mel*, em comparação com o próprio significado proposto para **Moacir**, o autor apresenta a definição "Filho do sofrimento: de *moacy*-dor, e *ira* – desinência que significa – *saído de*" (ALENCAR, 1965, p.148), sem a preocupação em indicar a língua a que pertence. A observação das explicações chama a atenção para o fato de que *ira* é, ao mesmo tempo, *mel* e a desinência *saído de*.

Entre os dicionários consultados, GD define "**Moacy** – magoar-se, estimular-se; agravado, sentido, doente" (p. 445) sem mencionar a etimologia. O dicionário apresenta definição semelhante para **Moacir**, porém não menciona a decomposição do vocábulo. Apesar de semelhante, o significado não é exatamente como Alencar propõe "o nascido do meu sofrimento". Sendo assim, **Moacir**, como **Iracema,** são considerados etimologia ficcional contextual de Alencar.

Outro antropônimo em que Alencar faz uso de *ira* no sentido de *mel* é em **Irapuã**. O autor traz, em nota, a explicação para o vocábulo, no romance *Iracema*

> **Irapuã** — De *ira* — mel, e *apuam* — redondo; é o nome dado a uma abelha virulenta e brava, por causa da forma redonda de sua colmeia. Por corrupção reduziu-se esse nome atualmente a arapuá. (ALENCAR, 1965, p. 150)

Na comparação com GD que apresenta o verbete "**Irapuã** – s. O cortiço redondo. De *ira*, abelha; *puã*, redondo. Riacho no R. G. do Sul. Cidade de São Paulo" (p. 551). Alencar desejava atribuir ao personagem características impressas no nome **Irapuã** que é ser bravo, hostil, vingativo e desejoso de vingança. Esse comportamento do índio **Irapuã** vai se construindo na narrativa, à medida que o narrador vai propondo o percurso de **Iracema**, por quem ele é apaixonado, porém a índia não poderá desposá-lo. Em razão disso, **Irapuã** vai se construindo um personagem com características negativas.

Com o propósito de continuar a investigação sobre o vocábulo **Irapuã**, recorremos ao *Corpus do Português* (DAVIES, 2016), que nos mostrou que, no século XIX, das 48 ocorrências de **Irapuã**, as 44 primeiras estão no romance *Iracema* de José de Alencar, as demais em Machado de Assis, em cujo texto tece críticas sobre o romance *Iracema*, portanto

trata-se de um texto posterior. A Figura 43 retrata as oito primeiras linhas de concordância do vocábulo **Irapuã** no *Corpus do Português* (DAVIES, 2016).

Figura 43: Linhas de concordância do vocábulo **Irapuã**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Considerando a análise do vocábulo **Irapuã**, podemos afirmar que Alencar cria uma etimologia ficcional contextual ao expandir o significado da segunda parte do vocábulo *apuam*, constante em GD para o nome de uma espécie de abelha brava e virulenta e justifica que é por causa da forma redonda de sua colmeia. Porém a pretensão na obra é aproximar o significado de **Irapuã** às características que desejou atribuir ao personagem com o mesmo nome.

Outro antropônimo empregado por Alencar é **Coatiabo.** Alencar o empregou para nomear **Martim**, após o "batizado" como membro da tribo dos pitiguaras. Conforme explicita Schwamborn (1998), no romance, **Iracema** escolhe o nome para **Martim** ao vê-lo pintado como o costume de sua tribo. O trecho a seguir de *Iracema*, retrata o momento do "batizado" de **Martim** na tribo dos pitiguaras:

> — Meu irmão é um grande guerreiro da nação pitiguara; ele precisa de um nome na língua de sua nação.
> — O nome de teu irmão está em seu corpo, onde o pôs tua mão.
> — Coatiabo! Exclamou Iracema.
> — Tu disseste; eu sou o guerreiro pintado; o guerreiro da esposa e do amigo. (ALENCAR, 1965, p. 112)

Neste trecho, Alencar já esboça o significado pretendido para o novo nome criado para **Martim** o "guerreiro pintado". Antes desse momento, porém, o autor já havia adiantado que era costume dos pitiguaras pintarem seus corpos com listras negras à semelhança do pelo do **quati**.

> Foi costume da raça, filha de Tupã, que o guerreiro trouxesse no corpo as côres de sua nação. Traçavam em princípio negras riscas sobre o corpo, à semelhança do pêlo do quati de onde procedeu o nome dessa arte da pintura guerreira. (ALENCAR, 1965, p. 112-113).

Em outra nota ao romance *Iracema*, Alencar explica que "*Coatiá* significa pintado. A desinência *abo* significa o objeto que sofreu a ação do verbo e, sem dúvida, provém de *aba*- gente, criatura". (p. 158). Então Alencar cria dois vocábulos: primeiro, *coatiá* - o verbo que indica ação de pintar e, depois, **Coatiabo** a partir de *coatiá* e *aba* – gente. Demonstra, portanto, que **Coatiabo** é, então, o guerreiro pintado.

Para corroborar nossa análise, fizemos a busca no *Corpus do Português* (DAVIES, 2016) e, as seis únicas ocorrências em todo o *corpus* foi no romance *Iracema* de Alencar, conforme Figura 44.

Figura 44: Linhas de concordância do vocábulo **Coatiabo**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Sendo assim, **Coatiabo** é um étimo de Alencar que traça todo um procedimento de criação do vocábulo e o explica em notas para que não pareça estranho aos olhos do leitor, portanto uma etimologia ficcional contextual. Schwamborn (1998) afirma que a cena e o

nome **Coatiabo** caracterizam a entrega por completo do europeu Martin à nova terra, o que se poderia assemelhar-se à naturalização como brasileiro.

Merece destaque também o vocábulo **Maranguab** que é empregado por Alencar como antropônimo e topônimo, conforme Figura 45, extraída por meio da ferramenta *Concord* do *WST*. É possível observar que, nas linhas um, dois e três, Alencar utiliza o vocábulo **Maranguab** para nomear uma serra, porém, nas demais, utiliza para nomear o índio, pai de **Jatobá**.

Figura 45: Linhas de concordância do vocábulo **Maranguab**

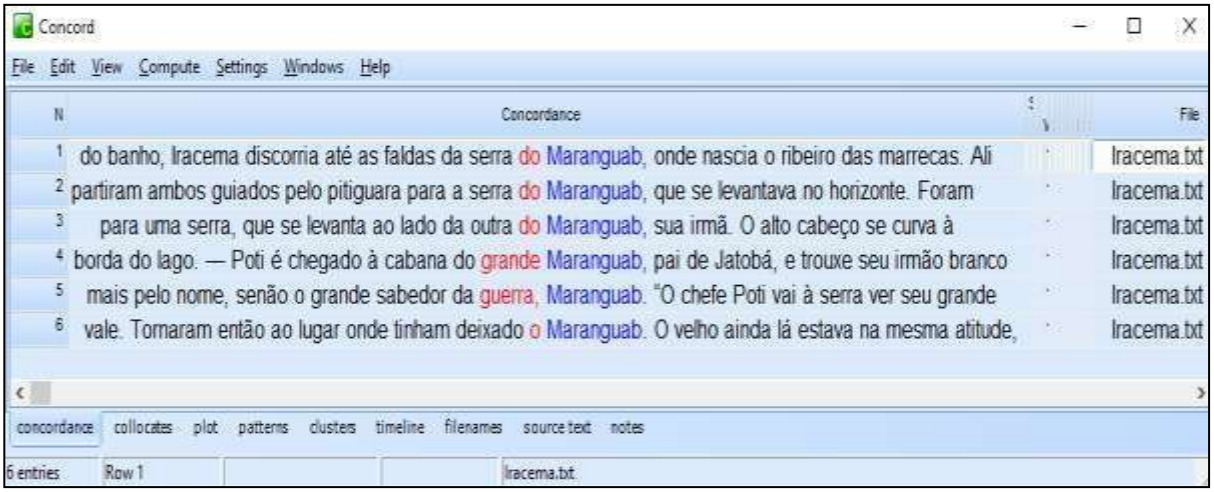


Figura: A autora, por meio da ferramenta *Concord* do *WST*

Em nota ao romance *Iracema*, Alencar explica

> **Maranguab** — A serra de Maranguape, distante cinco léguas da capital, e notável pela sua fertilidade e formosura. O nome indígena compõe-se de *maran* — guerrear, e *coaub* — sabedor; *maran* talvez seja abreviação de maramonhang — fazer guerra, se não é, como eu penso, o substantivo simples guerra, de que se fez o verbo composto. (ALENCAR, 1965, p. 157).

Nesta explicação, Alencar não menciona o fato de nomear também o índio guerreiro, porém, no corpo do romance *Iracema*, ele explica o que propõe com a escolha "Assim as tribos não o chamam mais pelo nome, senão o grande sabedor da guerra, Maranguab", conforme linha 5 da Figura 45, que justifica a explicação para o processo de formação *maran* – guerrear e *coaub* – sabedor, ou seja, o sabedor da guerra. Procedimento para justificar o nome pelo qual os pitiguaras chamavam **Batuireté**, **Maranguab**.

É possível observar o processo criativo de Alencar também a partir da continuação da nota sobre **Maranguab**, quando explica que o

> Dr. Martius traz etimologia diversa. *Mara* — árvore, *angai* — de nenhuma maneira, *guabe* — comer. Esta etimologia nem me parece própria ao objeto, que é uma serra, nem conforme com os preceitos da língua (ALENCAR, 1965, p. 157).

Nesse caso, ele demonstra que estudou a língua indígena, a ponto de discordar da formação proposta por outro autor e, portanto, cria uma etimologia ficcional contextual, pois ele próprio decompõe e compõe seus nomes e lhes atribui significado e justifica o processo de formação dos étimos.

Em relação ao emprego do vocábulo **Maranguab** como topônimo, Alencar o utiliza para denominar uma serra fértil e formosa que faz parte da região de **Maranguape**.

O vocábulo **Maranguab** não consta dos dicionários de consulta, porém, no *Corpus do Português* (DAVIES, 2016), as seis únicas vezes que este vocábulo é utilizado em textos compilados pelos autores é empregado por Alencar, em *Iracema*. Assim sendo, as linhas de concordância do *corpus* de estudos com as do *Corpus do Português* são exatamente as mesmas. A Figura 46, demonstra as linhas de concordância do vocábulo **Maranguab** no *Corpus do Português*.

Figura 46: Linhas de concordância do vocábulo **Maranguab**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

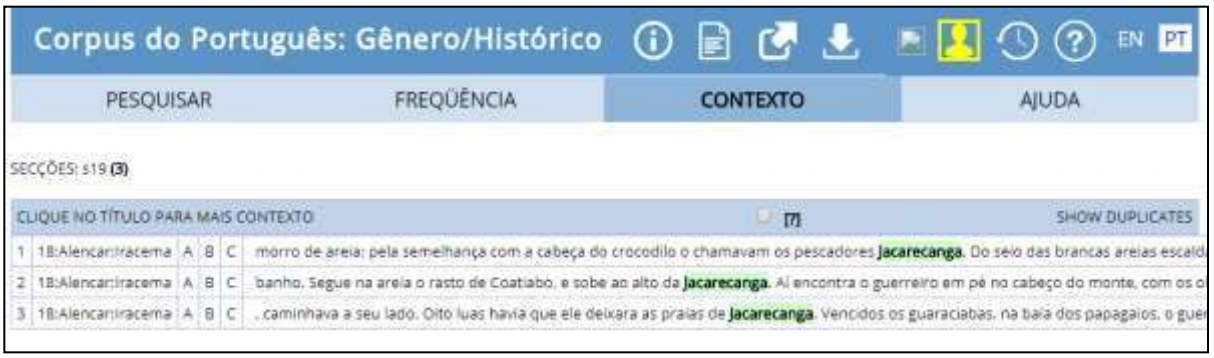
Ratificando, o étimo **Maranguab** é uma etimologia ficcional contextual de Alencar, tanto no emprego como antropônimo quanto topônimo, uma vez que não consta dos

dicionários de consulta e, no *Corpus do Português*, há apenas o emprego na obra *Iracema* de Alencar. Não se tem, de acordo com Davies (2016), outro emprego em textos escritos posteriores a Alencar. Neste caso, o vocábulo ficou restrito ao romance.

Outro exemplo de étimo de Alencar é **jacarecanga** que o autor assim o explica, em sua nota "**jacarecanga** – morro de areia na praia do Ceará, afamado pela fonte de água fresca puríssima. Vem o nome *jacaré* – crocodilo e *acanga* – cabeça" (ALENCAR, 1965, p. 159).

O vocábulo **jacarecanga** não consta dos dicionários de consulta, porém o vocábulo **jacaré** já havia sido dicionarizado por lexicógrafos anteriores à publicação de *Iracema*. Por exemplo, RB define "jacare', f. m. ou jacareo, (o primeiro mais comum no Brasil) o mesmo, que o crocodilo" (p. 740); e LMSP define "Jacare – s. m. Assim chamão no Brasil ao crocodilo" (s/d). Em textos escritos no Brasil do século XIX, a primeira ocorrência do vocábulo **jacarecanga** foi também por Alencar em *Iracema*, conforme Figura 47.

Figura 47: Linhas de concordância do vocábulo **jacarecanga**



Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Alencar teve a preocupação de explicar o significado do vocábulo **jacarecanga**, decompondo-o em nota "Morro de areia na praia do Ceará afamado pela fonte de água fresca puríssima. Vem o nome de *jacaré* — crocodilo, e *acanga* — cabeça" (p. 159), porque ele tinha consciência de que tratava de um vocábulo criado por ele, ou seja, uma etimologia ficcional contextual. O que também é comprovado pelo fato de que não consta o vocábulo **jacarecanga** em dicionários anteriores à publicação de *Iracema*, assim como o *Corpus do Português* (DAVIES, 2016) apresenta as únicas ocorrências, no século XIX, também em *Iracema*.

Há, ainda, outros antropônimos e topônimos que merecem destaque em análises, porém, como mencionado, não apresentaremos análises de todos, em razão da extensão da tese e da quantidade de vocábulos indígenas existentes no *corpus* de estudos. Numa primeira leitura, podemos nos deixar levar pela impressão de que Alencar é mais produtivo em relação aos étimos antroponímicos ou toponímicos, porém a criatividade se constata também em outros campos semânticos. Propomos, então, análises de vocábulos que não pertencem aos campos semânticos antropônimos e topônimos.

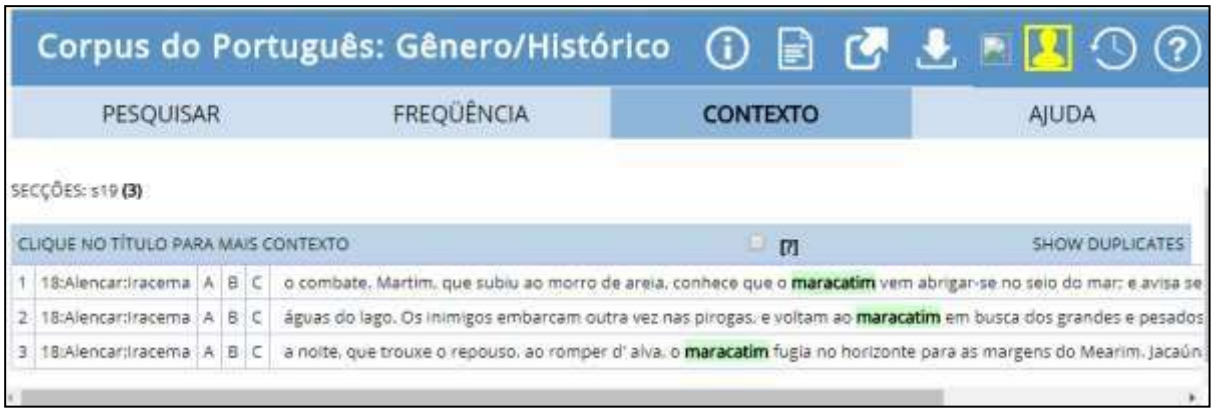
Ainda utilizando o processo de decomposição de vocábulos para depois compor outros, Alencar, em nota ao romance *Iracema*, baseado em Aires de Casal (1754?-1821?), decompõe o vocábulo **maracatim** em *maracá* – nome de um instrumento e *tim* – nariz, explicando que "Maracatim – grande barco que levava na proa um maracá" (ALENCAR, 1965, p. 159). O emprego do vocábulo na narrativa de *Iracema* pode ser considerado uma redundância, pois se o autor explica que **maracatim** é um "grande barco" e, ao mesmo tempo, o emprega acompanhado pelo adjetivo grande como na passagem "O grande maracatim corre nas ondas, ao longo da terra que se dilata até às margens do Parnaíba", torna-se, então, desnecessário a adjetivação.

Com base nos dicionários de exclusão, GD traz o verbete **maracatim** com a seguinte definição

> **Maracatim** – navio, embarcação grande. Era o nome que os índios davão às suas embarcações de guerra, as quaes tinhão na prôa um maracá, que eles fasião tocar quando acommettião" (GONÇALVES DIAS, 1858, p. 127)

A fim de investigar sobre a ocorrência do vocábulo em texto escritos no português brasileiro, recorremos ao *Corpus do Português* (DAVIES, 2016), conforme a Figura 48.

Figura 48: Linhas de concordância do vocábulo **maracatim**



Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Em relação ao vocábulo **maracatim**, observamos que em GD não há referências à sua constituição. Porém, Alencar estrutura uma decomposição para o vocábulo já dicionarizado, propondo-lhe uma etimologia ficcional contextual. Em textos escritos, no século XIX, apenas Alencar empregou o vocábulo **maracatim**, três vezes, no romance *Iracema*. Não foi possível identificar se se trata de um vocábulo já utilizado na língua portuguesa oral ou se fora utilizado em línguas indígenas ágrafas, porém o nosso *corpus* de consulta, o *Corpus do Português* (DAVIES, 2016), traz esse registro como utilizado pela primeira vez em textos escritos por Alencar, em *Iracema*.

Continuando o processo, Alencar cria étimos com base nas anotações realizadas, como também decompõe vocábulos a partir de outras já dicionarizadas na íntegra como para o vocábulo **igara**. Ele presume a decomposição para *ig* – água e *jára* – senhor; senhora das águas. GD informa que "*Ygára* – canoa. *Pocicába mondé Yara pupi*. Lastrar a canôa" (p. 505), portanto um significado diferente ao proposto por Alencar.

Como Alencar estudou os textos de GD, como constatado por Schwamborn (1998), ele, provavelmente, decompõe **igara** a partir da definição proposta por GD em outros verbetes como "*Jara* – Dono, amo, senhor, senhora" (p. 429) e "*Yg* – Agoa". (p. 504). Na nota ao romance, Alencar explica que "**igara**, de *ig* – água, e *jara* – senhor; senhora das águas". GD não explica a formação da palavra *Ygára* como o faz Alencar, portanto o sentido pretendido pelo autor de *Iracema* pode ser caracterizado como uma etimologia ficcional contextual.

Ainda sobre o vocábulo **igara**, pesquisamos no *Corpus do Português* (DAVIES, 2016) e há nove ocorrências para o vocábulo. Apesar de o primeiro texto referir-se a Guimarães, ele é posterior a Alencar, visto que se trata de Bernardo Guimarães. Esse dado foi obtido a partir da visualização das linhas de concordância do vocábulo. Ou seja, a primeira utilização do vocábulo em textos escritos é de José de Alencar, já que o texto "Histórias e Tradições da Província de Minas Gerais", em 1867, dois anos após a primeira edição de *Iracema*. A Figura 49, apresenta as linhas de concordância do vocábulo **igara**, para se constatar o exposto.

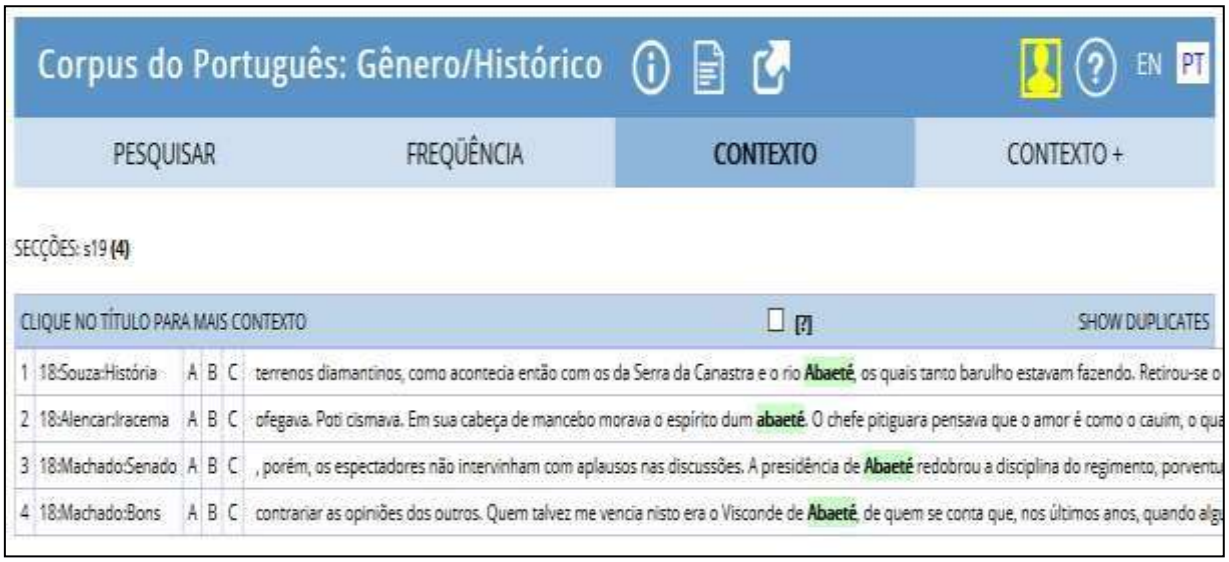
Figura 49: Linhas de concordância do vocábulo **igara**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Prosseguindo o estudo do léxico indianista do *corpus* de estudo de Alencar, procedemos à análise do vocábulo **abaeté**. Nos dicionários de língua portuguesa, RB, AMS e LMSP, anteriores à publicação dos romances de Alencar, não constam um verbete para **abaeté**. GD traz o verbete para **abaeté** com a seguinte definição "**Abaeté** s. Localidade de Minas Gerais. *Abá*, homem; *eté*, verdadeiro. Nome de um rio em Minas Gerais" (p. 513). Alencar traz abaeté, em nota, "**Abaeté** — Varão abalizado; de *aba* — homem, e *eté* — forte, egrégio" (ALENCAR, 1965, p. 155).

Utilizamos o *Corpus do Português* (DAVIES, 2016), para consulta, do qual extraímos as prováveis ocorrências da palavra em textos escritos no Brasil, conforme compilação dos autores. Como demonstrado na Figura 50, **abaeté** aparece 4 vezes no século XIX.

Figura 50: Linhas de concordância do vocábulo **abaeté**



Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Ao analisar a palavra abrimos as linhas de concordância e verificamos que **abaeté**, aparece em obra de Souza antes de *Iracema,* de Alencar. Realizamos uma busca no contexto do vocábulo ainda no *Corpus do Português* e encontramos a informação de que o autor seria Joaquim Norberto de Souza e Silva. O texto em que o vocábulo **abaeté** aparece trata-se da "História da Conjuração Mineira", publicado em 1821, anterior à publicação de Iracema. O autor cita **abaeté** no período

> Noticiou-lhe a vinda do Rio de duas companhias de tropa paga por ser pouca a que existia na capitania, em consequência de novos descobrimentos de terrenos diamantinos, como acontecia então com os da Serra da Canastra e o rio Abaeté, os quais tanto barulho estavam fazendo. (SOUZA e SILVA, 1821, s/p)[46]

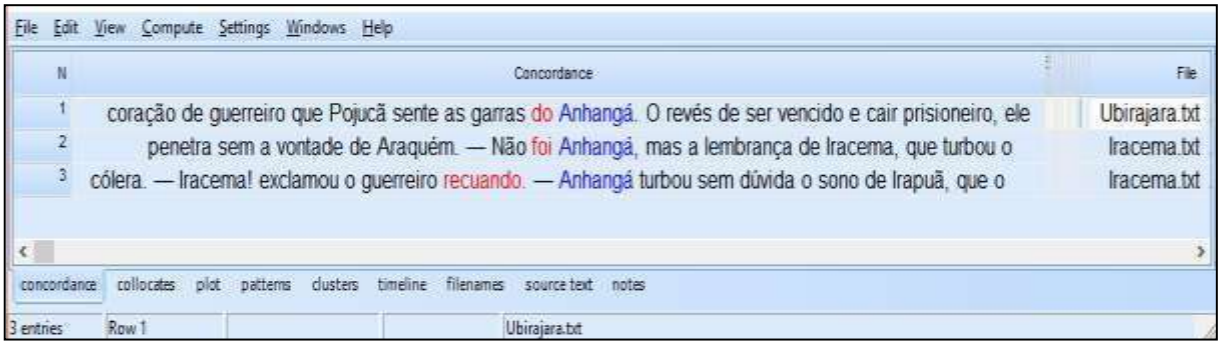
A partir desses dados não se pode afirmar que Alencar foi o primeiro a utilizar a palavra *abaeté* em textos escritos no Brasil, porém Souza o utiliza como o nome de um rio, enquanto Alencar utiliza referindo-se a um homem cujo espírito é bom. GD propõe a

[46] Trecho transcrito do *Corpus do Português* (DAVIES, 2016).

decomposição da palavra, porém explica que se trata de um rio. Alencar traz a nota para realçar as características desejadas na narrativa e explica o significado da palavra "abaeté – Varão abalizado; de *aba* – homem, e *etê* – forte, egrégio. (p. 155). Assim sendo, embora Alencar não seja o primeiro a empregar o vocábulo, trata-se de uma etimologia ficcional contextual porque o autor constrói um significado e o emprega divergindo dos dicionários e do texto de Sousa e Silva (1821).

Vale destaque também o vocábulo **anhangá** que é empregado três vezes no *corpus* de estudo, sendo duas vezes empregado em *Iracema* e uma vez em *Ubirajara*, conforme Figura 51.

Figura 51: Linhas de concordância do vocábulo **anhangá**



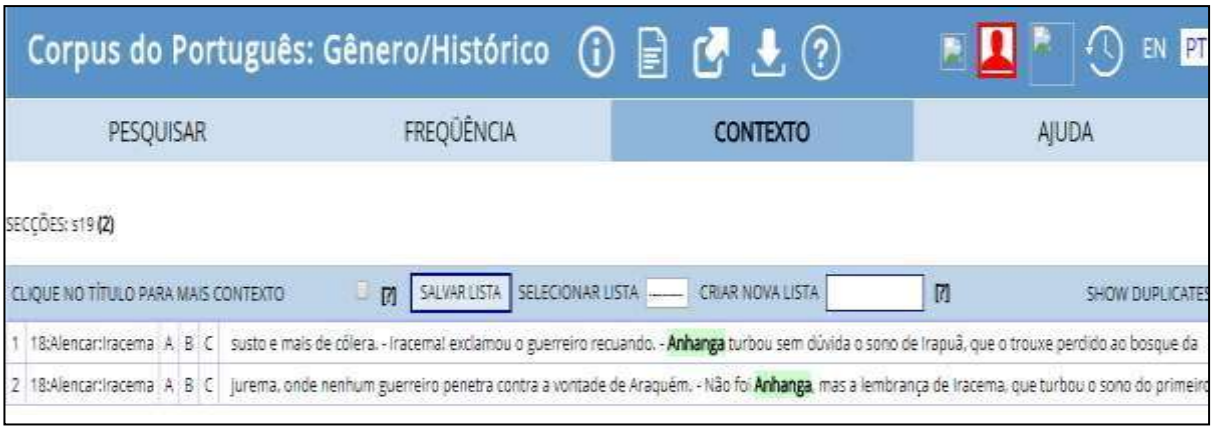
Fonte: A autora, por meio da ferramenta *Concord* do *WST*

Quando buscamos informações sobre o vocábulo no *Corpus do Português* (DAVIES, 2016), encontramos dois vocábulos: um com acento **anhangá** e outro sem acento **anhanga**. As Figuras 52 e 53 mostram, respectivamente as ocorrências.

Figura 52: Linhas de concordância do vocábulo **anhangá**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Figura 53: Linhas de concordância do vocábulo **anhanga**



Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Podemos notar que **anhanga**, sem acento, é empregado, por Alencar, em *Iracema*, enquanto **anhangá**, com acento, é empregado em *Ubirajara*. Porém, a comparação dos vocábulos, por meio das linhas de concordância, demonstra que se trata de uma questão de edição dos livros do autor, uma vez que os dois empregos se referem ao espírito. O próprio Alencar, em nota ao romance Iracema, explica "**Anhangá** — Davam os indígenas este nome ao espírito do mal; compõe-se de *anho* — só, e *angá* — alma. Espírito só, privado de corpo, fantasma". (ALENCAR, 1965, p. 152).

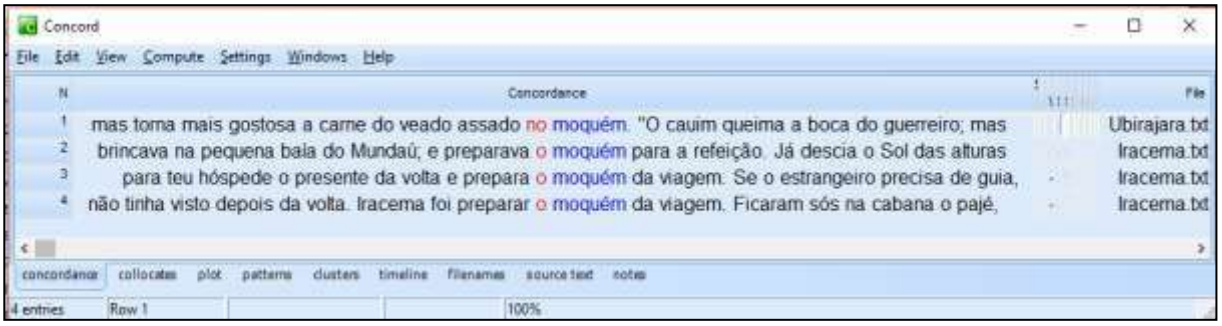
Entre os dicionários consultados, nenhum traz verbete para **Anhanga** ou **anhangá**. GD, porém, traz o vocábulo "**Anhangoara** – s. de *anhanga*, diabo; *goara, coara*, caverna. Localidade turística de S. Paulo: A caverna do diabo, série de grutas ainda não de todo exploradas". (p. 515).

Alencar, então, com base em GD, utiliza **anhanga** ou **anhangá** como espírito do mal, em extensão ao fato de que, para ele, seria o espírito sem corpo. Então, para a explicação Alencar decompõe o vocábulo a fim de explicar o emprego, portanto cria uma etimologia ficcional contextual para o vocábulo.

O vocábulo **moquém** foi analisado em Ávila (2004) como "tipo de comida indígena, que também era carregada em viagens", porém um estudo mais aprofundado da origem e do emprego desse vocábulo nos levou a uma análise distinta. Essa percepção foi motivada pela análise do vocábulo em *Ubirajara,* no qual deixa pistas mais precisas do significado do vocábulo. Na Figura 54, é possível notar as linhas de concordância em que o vocábulo

moquém é empregado em *Iracema* e *Ubirajara*. Na linha um está claro que se trata de um lugar onde se assam as carnes, porém nas demais linhas em que o autor emprega em *Iracema*, não deixa claro se se trata de uma comida ou lugar onde se prepara a comida.

Figura 54: Linhas de concordância do vocábulo **moquém**



Fonte: A autora, extraído por meio da ferramenta *Concord* do *WST*

A fim de dirimir a dúvida em relação ao emprego em *Iracema*, inicialmente, em consulta aos dicionários, notamos que GD traz um verbete para o vocábulo e assim o define "**Moquém** – s. Vila do Estado de Goiás. De *mo-caê*, fazer assar, espécie de grelha de varas para assar peixe, carne. Var. Muquem". (p. 571). Buscamos também as notas do autor em seus romances. Em *Iracema*, Alencar explica "**Moquém** — Do verbo *mocáem* — assar na labareda. Era a maneira por que os indígenas conservavam a caça para não apodrecer, quando a levavam em viagem. Nas cabanas a tinham no fumeiro" (ALENCAR, 1965, p. 152).

O vocábulo é explicado por processos de decomposição diferentes em Alencar e GD. Alencar afirma ser do verbo *mocáem* que já significa assar na labareda, porém GD diz ser formado por *mo-caê* que é uma espécie de grelha de varas para assar que viria de "fazer assar".

Embora o *Corpus do Português* (DAVIES, 2016) apresente ocorrências do vocábulo **moquém** apenas nas duas obras de Alencar: *Iracema* e *Ubirajara*, há uma diferença de em relação à quantidade de ocorrências. Há quatro ocorrências no *corpus* de estudo e, no *Corpus do Português*, seis, como se constata na Figura 55.

Figura 55: Linhas de concordância do vocábulo **moquém**

Fonte: A autora, extraído da *Corpus do Português* (DAVIES, 2016)

Em razão desta diferença em relação à quantidade de ocorrências, buscamos ampliar o contexto de abonação do vocábulo no *Corpus do Português* em relação ao livro *Ubirajara* e constatamos que se trata de uma nota do autor. Como as notas foram retiradas do *corpus* de estudo, justifica-se, então, a diferença em relação às ocorrências. Notamos também que Alencar explica em *Ubirajara* que "**Moquém** era simplesmente o assado envolto em folha e feito sobre a brasa", conforme Figura 56.

Figura 56: Contexto ampliado do vocábulo **moquém**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Não resta dúvida de que Alencar emprega **moquém** tanto para designar o processo de assamento ou o local onde se assam as carnes, como na definição de GD e sua nota, porém

utiliza também como o alimento, provavelmente, carnes que passam pelo processo, criando uma etimologia ficcional contextual para **moquém** no sentido de comida.

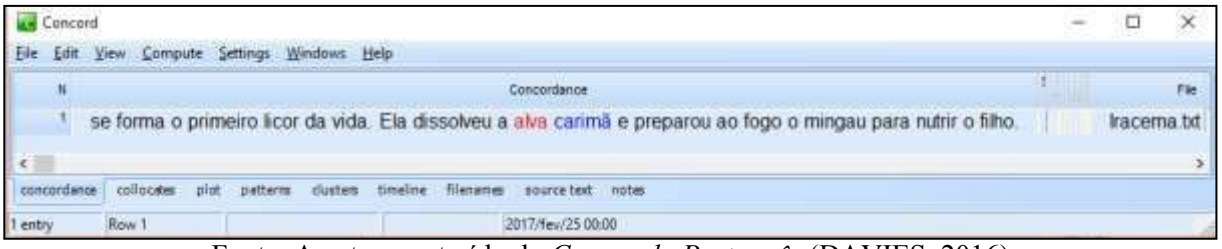
Outro vocábulo relacionado ao campo semântico alimentos é **carimã,** que é um vocábulo criado e utilizado pela primeira vez em textos escritos em Língua Portuguesa por Alencar. Isso pode ser comprovado pelo *Corpus do Português* (DAVIES, 2016), em que as duas ocorrências no século XIX estão em Alencar. A Figura 57 nos mostra as duas ocorrências de **carimã**, ambas em textos de Alencar, sendo a primeira em *Iracema*.

Figura 57: Linhas de concordância do vocábulo **carimã**

Fonte 1: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Linhas de concordância do vocábulo **carimã** Alencar emprega o vocábulo uma única vez, no *corpus* de estudos, como se constata com a linha de concordância obtida pela ferramenta *Concord* do *WST*, de acordo com a Figura 58.

Figura 58: Linhas de concordância do vocábulo **carimã**



Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Além disso, os dicionários consultados GD, LMSP, RB e AMS não trazem o vocábulo **carimã** entre seus verbetes. Talvez, por isso, Alencar teve a preocupação, pelo fato de não ser de conhecimento dos leitores, de trazer uma nota explicando o vocábulo "**Carimã** — Uma conhecida preparação de mandioca. *Caric* — correr, *mani* — mandioca: mandioca escorrida". (AELNCAR, 1965, p. 160).

Note-se, então, que Alencar cria uma etimologia ficcional contextual ao explicar o significado do vocábulo, por meio da decomposição *caric* – correr + *mani* – mandioca.

Outros dois vocábulos destacados do *corpus* de estudo são **camucim** e **camuci**. Em relação ao vocábulo **camucim**, Alencar também demonstra um processo criativo ao utilizá-lo em suas obras. Em nota ao romance *Iracema*, o autor assim o explica

> **Camucim** — Vaso onde encerravam os indígenas os corpos dos mortos e lhes servia de túmulo; outros dizem *camotim*, e talvez com melhor ortografia, porque, se não me engano, o nome é corrupção da frase *co* — buraco, *ambira* — defunto, *anhotim* — enterrar; buraco para enterrar o defunto: *c´ am´ otim*. O nome dava-se também a qualquer pote. (ALENCAR, 1965, p. 152).

Já GD define "**Camucim** – Cidade do Ceará, o mesmo que cambucy. (p. 528); e **Cambucy** em outro verbete "**Cambucy** – apresenta as variantes camucym, camotim, significando vaso, pote. Nome do bairro da capital paulista, vila dos Estados do Rio de Janeiro" (p. 527).

Com base nas linhas de concordância, conforme Figura 59, é possível perceber que Alencar emprega o vocábulo **camucim** em três acepções. Nas linhas um, três, quatro e doze é empregado como antropônimo o que revela o emprego de um vocábulo ressignficado. Já nas outras linhas, tem relação com o explicado por Alencar e definido por GD que nomeia um tipo de féretro em que os índios eram colocados para o sepultamento ou um tipo qualquer de pote.

Nas linhas dois, sete, oito e nove, ambas as ocorrências de *Ubirajara*, narra a cena em que **Jurandir**, um guerreiro indígena, deverá mostrar sua bravura introduzindo sua mão dentro de um formigueiro de saúvas, uma das espécies de formigas muito venenosas. Esse ritual demonstrava a força e coragem do guerreiro que submetia a sua própria mão às picadas das formigas. Fato que, na linha sete, é narrado que o vaso teve que ser quebrado, porque a mão de **Jurandir** já não passava mais pelo buraco do pote, em razão do inchaço intenso e da inflamação que adquiriu com o veneno das formigas.

Já nas linhas 11 e 14, o autor faz referência a um tipo de pote pequeno, onde se guardavam umas castanhas com as quais se contavam a idade dos índios. O procedimento era

de guardar nesse pote uma castanha para cada florada do caju, o que marcava, cada castanha, um ano de vida o índio.

Nas linhas cinco, seis, dez e treze o autor utiliza **camucim** como um tipo de féretro onde os índios colocavam os corpos dos seus entres para o sepultamento.

Figura 59: Linhas de concordância do vocábulo **camucim**

Fonte: A autora, por meio da ferramenta *Concord* do *WST*

A fim de verificar informações sobre o emprego do vocábulo **camucim** em outros textos escritos em Língua Portuguesa, consultamos o *Corpus do Português* (DAVIES, 2016), o qual tem registrado que, no século XIX, há apenas as ocorrências nas obras *Iracema* e *Ubirajara* de José de Alencar, A Figura 60 demonstra as linhas de concordância do vocábulo **camucim**.

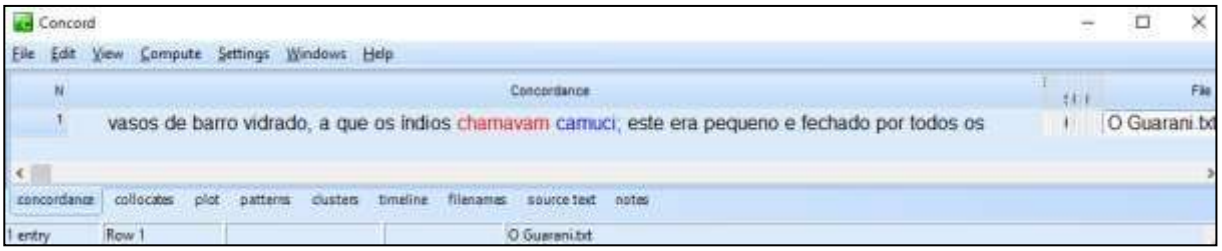
Figura 60: Linhas de concordância do vocábulo **camucim**

Fonte: A autora, extraído do *Corpus do Português* (DAVIES, 2016)

Fica evidente, portanto, que Alencar explora o vocábulo **camucim** em diversas possibilidades, ou seja, emprega para nomear um rio, um féretro ou, simplesmente, um tipo de pote utilizado para guardar sementes ou outro tipo de receptáculo para criar um formigueiro de saúvas. Trata-se, portanto, de uma etimologia ficcional contextual.

Outro aspecto a ser observado foi em relação ao vocábulo **camuci** utilizado em *O Guarani*, pois ele apresenta as mesmas características de **camucim**. É possível observar na Figura 61 que também é um tipo de pote de barro utilizado para diversos fins pelos índios.

Figura 61: Linhas de concordância do vocábulo **camuci**



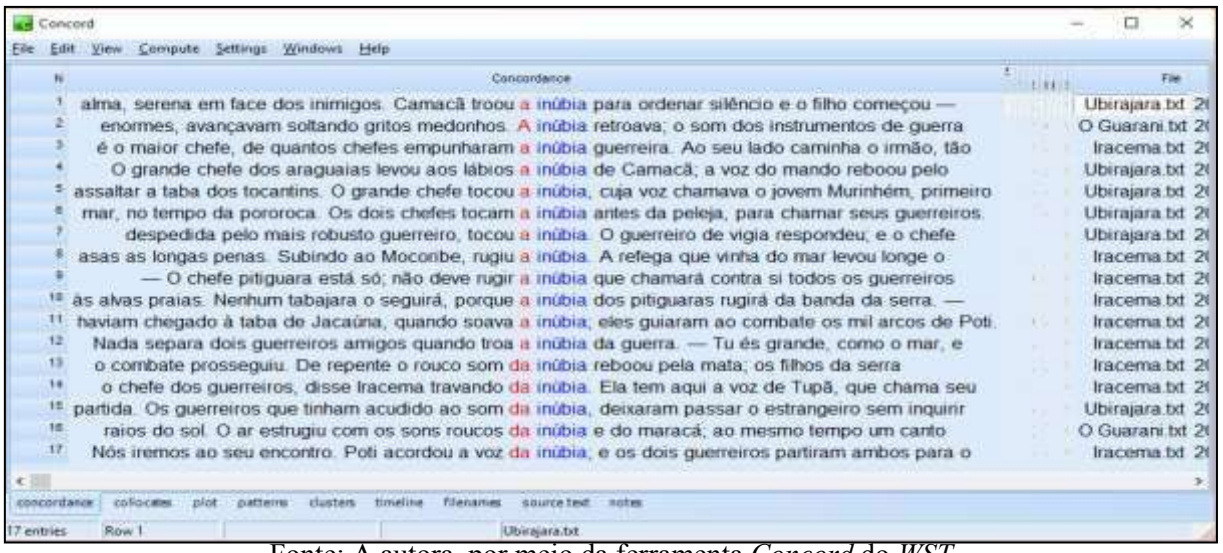
Fonte: A autora, por meio da ferramenta *Concord* do WST

No *Corpus do Português*, o vocábulo **camuci** aparece uma única vez em todo o *corpus,* no texto *O Guarani* de Alencar, reproduzindo a mesma linha de concordância

apresentada pelo *WST*. Trata-se, portanto, de um étimo de Alencar, porque também não consta nos dicionários de consulta.

Outro vocábulo utilizado por Alencar que merece destaque é **inúbia**. Ele não apresenta discrepância em relação ao processo de formação, contudo Alencar, em nota, explica que é uma espécie de trombeta de guerra utilizada pelos índios. Em todas as ocorrências o vocábulo é utilizado em contexto referente a batalhas entre os índios. A Figura 62 apresenta as 17 ocorrências nos romances de Alencar. Ressalta-se que o vocábulo é utilizado nas três obras que compõem o *corpus* de estudo, igualmente como um instrumento musical que é tocado, associado à guerra.

Figura 62: Linhas de concordância do vocábulo **inúbia**



Fonte: A autora, por meio da ferramenta *Concord* do *WST*

Buscando o vocábulo no *Corpus do Português* (DAVIES, 2016), encontramos as mesmas linhas de concordância apresentada pelo *WST*, ou seja, no século XIX, Alencar foi o único a empregar este vocábulo. Ressalta-se que nos baseamos nas informações contidas a partir dos textos compilados e disponibilizados pelo *Corpus do Português*. Em razão de as linhas de concordância serem as mesmas no *corpus* de estudo e no *Corpus do Português*, julgamos desnecessário reproduzir as linhas de concordância do *Corpus do Português*.

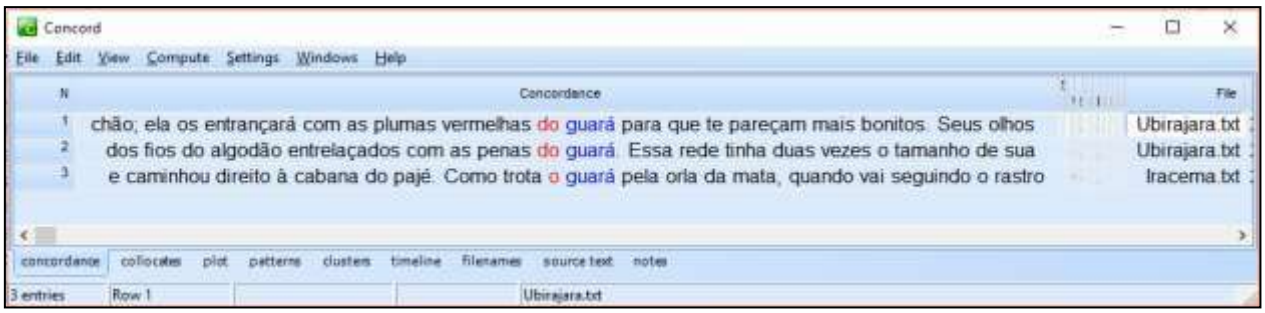
Considerando também que GD e demais dicionários anteriores a Alencar não trazem o vocábulo **inúbia** entre seus verbetes, podemos afirmar que se trata de uma etimologia ficcional contextual de Alencar.

Trazemos mais outros dois vocábulos empregados por Alencar que merecem análise: **guará** e **gará**. Iniciamos com a explicação que o autor atribui ao vocábulo **guará,** em nota ao romance *Iracema,* "**Guará** — Cão selvagem, lobo brasileiro. Provém esta palavra do verbo **u** — comer, do qual se forma com o relativo **g** e a desinência **ara** o verbal **g-u-ára** — comedor. A sílaba final longa é a partícula propositiva **ã** que serve para dar força à palavra" (p. 153).

No cotejo com os dicionários, verificamos a discrepância com GD em que **guará** é definido como "**Guará** – ave. Nasce branca, torna-se preta e por fim de um escarnado vivíssimo. (p. 418). Na seção do dicionário de GD destinada aos topônimos, o autor define também **guará** como "Ribeiro da Bahia. De *guará*, garça". (GONÇALVES DIAS, p. 539).

Observando as linhas de concordância do *corpus* de estudo, podemos perceber que Alencar utiliza **guará** como ave no romance *Ubirajara*, conforme as linhas um e dois da Figura 63. Já no romance *Iracema*, Alencar utiliza como um animal carnívoro à procura de sua presa.

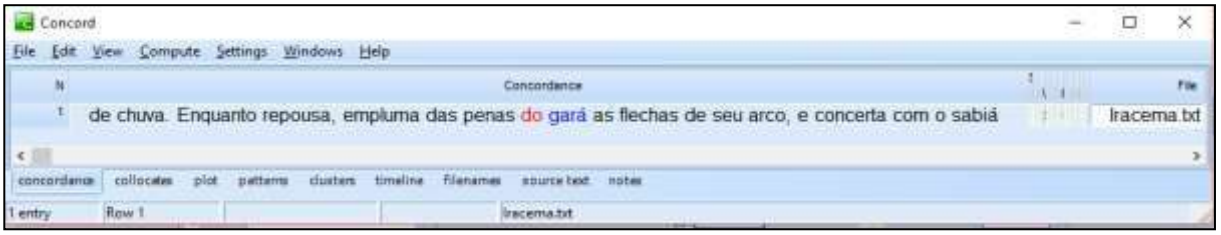
Figura 63: Linhas de concordância do vocábulo **guará**



Fonte: A autora, por meio da ferramenta *concord* do *WST*

Pesquisamos o vocábulo **guará** no *Corpus do Português* (DAVIES, 2016) e as linhas de concordância são exatamente as mesmas do *corpus* de estudo, por isso nos abstivemos de trazer a figura extraída do *Corpus do Português*. Verificamos também as linhas de concordância do vocábulo **gará** e obtivemos os dados, conforme a Figura 64.

Figura 64: Linhas de concordância do vocábulo **gará**



Fonte: A autora, por meio da ferramenta *Concord* do *WST*

Nota-se que a ave **gará,** em *Iracema,* tem o mesmo valor do vocábulo **guará** de *Ubirajara* que é referente a uma espécie de ave de penas vermelhas. Dado que é corroborado pela nota do autor, em Iracema "**Gará** — Ave paludal, muito conhecida pelo nome de guará. Penso eu que esse nome anda corrompido de sua verdadeira origem, que é *ig* — água, e *ará* — arara: arara d´água. Também assim chamada pela bela cor vermelha". (ALENCAR, 1965, p. 148).

De acordo com os estudos de Schwamborn (1998), Alencar buscou em Aires de Casal (1754?-1821?), informações sobre o vocábulo **guará**, porém Alencar contesta a explicação de Aires de Casal justificando por meio de longa explicação sobre os hábitos e característica dessa ave.

Então se Alencar próprio oferece explicação para **guará** sendo uma ave, ainda assim utiliza o mesmo vocábulo para indicar uma espécie de lobo, um animal feroz. Nesse caso, o autor propõe uma decomposição para o vocábulo criando uma etimologia ficcional contextual para o **guará,** empregado como animal, e **gará,** empregado como ave.

Podemos presumir que Alencar tenha descoberto uma forma ou um padrão para decompor as palavras indígenas em elementos significativos e, em seguida, compor novos vocábulos de acordo com suas necessidades ficcionais no contexto das obras escritas.

Era comum a nomeação de um lugar motivado por fatores extralinguísticos. A representação dos espaços geográficos tem relação com fatores de ordem histórica, social e cognitiva, ou seja, no ato do batismo de uma localidade havia uma tendência a se basear nos fatores que a caracterizariam. Alencar manteve alguns topônimos como já utilizados pelos autores estudados como Gonçalves Dias, Aires de Casal e Varnhagen. Assim, Alencar foi estruturando seus romances, ora com vocábulos já existentes, ora os ressignificando. Em

outros momentos, a partir de sua própria decomposição, cria novos étimos com os segmentos decompostos.

Ao realizar as análises buscamos as edições atuais dos livros de Aires de Casal e Varnhagen para verificar as fontes nas quais Alencar se baseou para criar seus étimos indianistas. Encontramos, a título de exemplo, o vocábulo **caiçara,** que Aires de Casal (1976, p. 139)[47] traz o enunciado "Defronte está a Real Fazenda da Caiçara, onde se cria numeroso gado vacum, e também cavalar, e onde se tem morto grande número de tigres". Há menção ao vocábulo caiçara, porém o autor não explica a sua formação, nem atribui significado a ele. Em GD, há apenas a menção "Cayçara – s. Trincheira, arraial" (p. 398), ou seja, também não explica a formação do vocábulo. Alencar, então, com base na decomposição de outros vocábulos também o faz para caiçara e propõe "**Caiçara** — De *cai* — pau queimado e a desinência *çara*, cousa que tem, ou se faz; o que se faz de pau queimado. Era uma forte estacada de pau-a-pique."

Outro exemplo é o vocábulo **aracati**. Aires de Casal (1976, p. 288) traz o vocábulo no seguinte enunciado "A Vila de São Bernardo, ainda pequena, está situada junto a uma ribeira, que depois de meia légua se ajunta ao Jaguaribe pela margem ocidental, obra de dez léguas acima de Aracati, e quarenta abaixo de Icó". Neste caso, não há também referência à formação ou ao significado do vocábulo. Alencar, porém, o decompõe e explica, em nota ao romance Iracema:

> **Aracati** — Significa este nome bom tempo — de *ara* e *catu*. Os selvagens do sertão assim chamavam as brisas do mar que sopram regularmente ao cair da tarde e, correndo pelo vale do Jaguaribe, se derramam pelo interior e refrigeram da calma abrasadora do verão. Daí resultou chamar-se Aracati o lugar de onde vinha a monção. Ainda hoje no Icó o nome é conservado à brisa da tarde, que sopra do mar. (ALENCAR, 1965, p. 152).

A decomposição de Alencar destoa da proposta por GD "De ara, vento; caty, maresia", o que comprova que Alencar se propõe a criar vocábulos indígenas a partir de suas pretensões literárias e linguísticas.

[47] Esta é uma edição atualizada daquela que Alencar utilizou, cuja publicação pode ter sido em 1754 ou 1821. Na edição de 1945, o trecho vem escrito com a seguinte grafia "Defronte está a Real Fazenda da *Cahyssára*, onde se cria numerozo gado vacum, e tambem cavalar; e onde se tem morto grande numero de tygres." (p. 301) (destaque do autor).

Os exemplos desse processo criativo se multiplicam se comparar os livros de Aires de Casal, Varnhagen e GD com as obras de Alencar, porém nos ateremos aos dois como critério de exemplificação, já que não é propósito desta tese estabelecer estas comparações, porém elas ratificam as análises propostas.

Assim sendo, Alencar vai estabelecendo sua etimologia ficcional contextual na medida em que deseja aproximar o significado da palavra com o elemento nomeado. A decomposição de nomes já existentes serve para interpretar, principalmente, os topônimos já existentes, porém sem explicação. Por isso, a utilização de frases explicativas no próprio texto ou a recorrência a notas explicativas ao final dos romances.

Vale ressaltar que os étimos criados por Alencar não são aleatórios, existe, por trás de cada um, uma razão que é esclarecida no decorrer da narrativa, nas notas ou nos textos de defesa escritos após a publicação de cada romance. Alencar cuida de dar nomes aos elementos do mundo indígena de acordo com a etimologia de palavras de origem Tupi ou Guarani, dentre outras línguas indígenas. A maioria dos vocábulos indígenas são utilizados para atender aos desígnios do autor e cumprir seu objetivo de criação de uma língua brasileira, a partir da figura do indígena.

No capítulo seguinte, apresentamos uma proposta de vocabulário *online* com os vocábulos indígenas da trilogia indianista de Alencar: *Iracema*, *O Guarani* e *Ubirajara*.

CAPÍTULO 7 - LÉXICO INDIANISTA DE JOSÉ DE ALENCAR: PROPOSTA PARA UM VOCABULÁRIO *ONLINE*

Este capítulo tem por pretensão atender ao objetivo da tese que é propor um desenho para elaboração de um Vocabulário com os vocábulos indianistas de Alencar, a ser disponibilizado eletronicamente via *internet*. Sendo assim, optamos por expor uma amostra do vocabulário com os itens lexicais indianistas que integram a tríade de José de Alencar: *O Guarani*, *Iracema* e *Ubirajara*. Como se trata de uma proposta, não apresentamos fichas e verbetes de todos os vocábulos indígenas, mas daqueles que servem para esboçar a proposta do vocabulário *online* que buscaremos realizar, posteriormente, a partir desta pesquisa. Portanto, sem a pretensão de atingir a totalidade dos vocábulos indianistas, como mencionado, apresentamos 15 vocábulos que exemplificam o léxico indianista; entretanto, esses 15 vocábulos se desdobram em 17 fichas lexicográficas, tendo em vista que dois deles são empregados em duas acepções cada um.

Este capítulo abarca, portanto, as fichas lexicográficas de 17 vocábulos e os seus verbetes correspondentes; um breve apontamento que busca justificar a inserção de antropônimos e topônimos entre os verbetes do vocabulário e, por fim, a nossa proposta de elaboração do vocabulário indianista de Alencar. Isso porque, segundo Fromm (2007, p. 38), "um dos pontos básicos para a elaboração de um banco de dados é a criação de uma estrutura para organizar a informação a ser coletada"; sendo assim, a elaboração das fichas e dos verbetes nos fornecerão os dados que serão utilizados para a organização do vocabulário *online*.

7.1 Elaboração do vocabulário indianista

Para fins de elaboração de um vocabulário *online*, a organização dos dados, dentre eles, a elaboração da ficha lexicográfica que foi pautada no público potencial consulente, ou seja, leitores das obras, professores de literatura e discente em geral. Por se tratar de uma proposta com uma amostra dos vocábulos indianistas usados por Alencar, portanto não um produto da tríade como um todo, as formas de consulta são sugestões e poderão ser alteradas de acordo com o processamento do vocabulário *online*. A proposta é que os dados das fichas sejam informatizados para a geração do vocabulário em um ambiente de gestão *online*.

Consideramos pertinente a elaboração e publicação de um vocabulário, pois, ao realizarmos pesquisas na *Web,* não encontramos um documento que retrate o léxico indianista de José de Alencar, em específico. Encontramos, entretanto, um número extenso de trabalhos sobre o autor, em diversas abordagens de pesquisa, e alguns dicionários de língua indígena como o *Dicionário Ilustrado Tupi Guarani*[48], que apresenta um suporte de pesquisa por meio da escolha da palavra, por ordem alfabética ou por meio de um ícone "dicionário" que disponibiliza itens como bichos, culinária, nomes de lugares, nomes indígenas, plantas e povos nativos, com algumas referências aos contextos abonatórios. Encontramos também o *Dicio, Diconário on line de Português*, *Dicionário Indígena*, *Minidicionário indígena* e *Minidicionário Tupi-Guarani*, porém esses, com o título de dicionário, se atêm a listar as palavras com um significado desprovido de etimologia, processo de formação e contexto abonatório.

Em síntese, pretendemos apresentar uma obra de referência que acreditamos que poderá ser útil para os leitores e estudiosos de Alencar. O vocabulário pretende explicitar as nuances referentes ao léxico indianista de José de Alencar em *Iracema*, *O Guarani* e *Ubirajara* e poderá fornecer subsídios para auxiliar os consulentes na compreensão e interpretação desses romances, visando a conhecer as peculiaridades do autor, bem como expandir os conhecimentos sobre os romances.

Acreditamos, por fim, que, do ponto de vista dos estudos lexicais da língua portuguesa, este vocabulário poderá interessar pelas informações relacionadas à etimologia e formação das palavras indígenas que contribuíram para a formação da língua portuguesa brasileira.

Na seção seguinte, procedemos à exposição das fichas lexicográficas elaboradas.

[48] Disponível em www.dicionariotupiguarani.com.br. Acesso em: 02 maio 2018.

7.2 Fichas lexicográficas e verbetes: um recorte do léxico indianista

A estruturação das fichas lexicográficas é um componente essencial para a produção do vocabulário, já que elas agregam as informações necessárias para elaboração dos verbetes dos vocábulos e da organização do vocabulário a ser disponibilizado *online*.

Conforme relatado anteriormente, para cada um dos vocábulos construímos uma ficha lexicográfica, a qual nos fornece as informações imprescindíveis para a elaboração dos verbetes que comporão o vocabulário dos itens lexicais indianistas do autor. Para elaboração das fichas, consultamos os contextos abonatórios adquiridos, por meio da ferramenta *Concord* do *WST*, a partir dos romances indianistas de Alencar. Além disso, quatro dicionários de língua indígena foram consultados:

i) *Dicionário de língua Tupy***,** de A. Gonçalves Dias (1858) - **GD;**

ii) *O tupi da geografia nacional***,** de Theodoro Sampaio (1901**) – TS;**

iii) *Dicionário Histórico das palavras Portuguesas de origem Tupi***,** de Antônio Geraldo da Cunha (1998) – **AGC.**

iv) *O Dicionário Tupi-português*, de Luiz Caldas Tibiriçá (1984) – **LCT**

Para a elaboração de nossa ficha lexicográfica adotamos como base os procedimentos de produção de verbetes de Ávila (2004); Ávila e Martins (2008) e Schreiner (2012), que estão de organizadas de acordo com os seguintes critérios:[49]

i) enumeradas, como critério de organização;

ii) as fichas constam dos seguintes elementos em sua constituição:

- Entrada – que é o vocábulo;

- Identificação das obras em que o vocábulo é utilizado pelo autor;

- Identificação da classe gramatical;

- Identificação do campo semântico;

[49] Os detalhes sobre a formação das fichas foram descritos no capítulo de Metodologia.

- Definição contextual;

- Informações sobre a etimologia e o processo de formação;

- Passagens abonatórias;

- Definição dos dicionários consultados;

- Notas, quando necessário.

Para as unidades lexicais que possuem mais de um significado, serão elaboradas uma ficha para cada significado depreendido, como o do vocábulo **andira**, que, com base nos contextos abonatórios, possui dois empregos: como antropônimo e como elemento da fauna. Como se trata de uma proposta, apresentamos análise de 15 vocábulos, dos quais resultaram 17 fichas lexicográficas. As fichas foram elaboradas adotando como critério de selecionar os primeiros vocábulos da sequência dos campos semânticos: antropônimos, topônimos, etnônimos, flora, fauna, utensílios/vestimenta, alimentos, espiritualidade, habitação, diversos. Excetua-se **jatobá**, o qual, apesar de não estar entre os primeiros das listas, foi mantido. Isso porque ele foi utilizado, inicialmente, para testagem do estudo e, então, resolvemos mantê-lo.

A nossa proposta é estabelecer uma descrição que considere o valor dos vocábulos indianistas no contexto das obras, não fora deste. Consideramos como descrição contextual a definição de um item lexical pautada nos elementos que estão associados a ele. Nesse caso, não só as propriedades semânticas do item são pertinentes, mas todo o contexto que evolve aspectos sociais e culturais. No caso do indianismo em Alencar, a descrição contextual ou o significado do léxico indigenista adquire *status* dentro dos textos, portanto, para nós, é imprescindível o contexto linguístico e semântico de uso para descrever os vocábulos. Isso corrobora o que afirmou Barbosa (2014, p. 417): “definir é o processo de analisar e descrever o semema linguístico, para reconstruir o modelo mental: o seu ponto de partida é a estrutura linguística manifestada”.

Após cada ficha apresentamos, também, o verbete correspondente ao vocábulo. A seguir, as fichas preenchidas dos vocábulos: **abacaxi**, **abaré**, **abati**, **acajás**, **acaraú**, **acauã**, **aimoré**, **anajê**, **andira** (antropônimo), **andira** (fauna), a**racati**, **araçoia**, **araguaia**, **boré**, **itaoca**, **jatobá** (antropônimo), **jatobá** (flora).

Ficha Nº: 01		
Abacaxi	s.m.	Elemento da flora
Descrição contextual: Fruta de odor agradável da família das bromélias. Possui uma casca com a semelhança de gomos e, sobre a fruta, folhas com bordos espinhosos que lembram uma coroa.		
Etimologia e/ou Processo de formação: *Ibacaxi*, c. *ibá*. Fructa + *caxi*= cati, rescendente, cheirosa = fruta cheirosa		
Passagens abonatórias: Frutas de várias espécies, pencas douradas de bananas, cachos roxos de açaí, os rubros croás e os fragrantes <abacaxis>, enchiam o jirau levantado no meio do terreiro. (Ubirajara)		
Definição dos dicionários consultados: GD: n/c TS: corr. *Ibacaxi*, c. *ibá*. Fructa, *caxi*= cati, rescendente, cheirosa. (107) AGC: *s.m*. Var.: 8 *abacachí*, 8-9 *abacaxi*, 9 *abacaxí* [,T. iuaka'ti < ï'uia 'fruta' + ka'ti 'rescendente'. Fruto carnoso, comestível, de uma planta cultivada da família das bromeliáceas; variedade do ananás. (p. 41) LCT: espécie de ananás. (p. 47)		
Nota: Não há discrepância em relação às definições apresentadas pelos dicionários consultados, apenas uma distinção fonética da etimologia proposta por TS e AGC. Ressalta-se que o dicionário de GD anterior à publicação das obras de Alencar, não traz **abacaxi** entre os seus verbetes. Alencar utiliza o vocábulo **abacaxi** no mesmo sentido apresentado pelos dicionários e o emprega apenas uma vez no romance *Ubirajara*.		

Verbete:

Abacaxi. s.m. Elemento da flora. Etim. de *Ibacaxi*, c. *ibá*. Fructa + *caxi*= cati, rescendente, cheirosa ou *iuaka'ti* < *ï'uia* 'fruta' + *ka'ti* - rescendente. Definição contextual: Fruta de odor agradável da família das bromélias. Possui uma casca com a semelhança de gomos e, sobre a fruta, folhas com bordos espinhosos que lembram uma coroa. Passagens abonatórias: 1 *Frutas*

de várias espécies, pencas douradas de bananas, cachos roxos de açaí, os rubros croás e os fragrantes <abacaxis>, enchiam o jirau levantado no meio do terreiro. (Ubirajara)

<table>
<tr><td colspan="3">Ficha N^{o}: 2</td></tr>
<tr><td>Abaré</td><td>s. m.</td><td>Espiritualidade</td></tr>
<tr><td colspan="3">Definição contextual:

Pessoas de prestígio na tribo que possuíam um poder de decidir questões importantes. Elas eram consideradas como uma espécie de instrutores espirituais da tribo.</td></tr>
<tr><td colspan="3">Etimologia e/ou Processo de formação:

abá, home + ré, diferente = homem diferente</td></tr>
<tr><td colspan="3">Passagens abonatórias:

O conselho dos <abarés> se reunira para meditar sobre a guerra. O velho Majé, a quem irritava o desaparecimento da filha, reparou que sem o voto do carbeto se convocasse a nação. (Ubirajara)

Muitas outras coisas referiu Jurandir; e os anciões admiravam-se de ver o juízo prudente de um <abaré> no corpo jovem de tão forte guerreiro. (Ubirajara)</td></tr>
<tr><td colspan="3">Definição dos dicionários consultados:

GD: - s. Hoje Avaré, cidade do Est. De S. Paulo. De abá, home; ré, diferente, que não é igual aos outros, isto é, o padre. (p. 513)
TS: c. abá-ré, homem distincto, diferente dos outros, o padre ou missionário. (p.107)
AGC: s.m. Var.: 5 abarê, 5-8 abaré [<T. aua’re VLB Clerigo ou frade + Abarê] Designação que os indígenas davam aos padres, particularmente aos jesuítas.
LCT: sacerdote (p. 47)</td></tr>
<tr><td colspan="3">Nota: Os dicionários consultados trazem abaré referindo-se a membro da igreja católica, porém no romance Ubirajara não há contato entre índios e brancos, isto é, Alencar narra a história que finaliza com a união de duas tribos. Sendo assim, no contexto do romance, abaré não significa padre, mas refere-se àqueles que, à semelhança dos padres entre os católicos, têm a função de orientar espiritualmente os membros da tribo. É considerado como campo semântico espiritualidade em razão de serem guias espirituais da tribo.</td></tr>
</table>

Verbete:

Abaré. s.m. Espiritualidade. Etim. *abá*, home + *ré*, diferente. Definição contextual: Pessoas de prestígio na tribo que possuíam um poder de decidir questões importantes. Elas eram consideradas como uma espécie de instrutores espirituais da tribo. Passagens abonatórias: 1) *O conselho dos <abarés> se reunira para meditar sobre a guerra. O velho Majé, a quem irritava o desaparecimento da filha, reparou que sem o voto do carbeto se convocasse a nação. (Ubirajara). 2) Muitas outras coisas referiu Jurandir; e os anciões admiravam-se de ver o juízo prudente de um <abaré> no corpo jovem de tão forte guerreiro. (Ubirajara).*

Ficha Nº: 3		
Abati	s. m.	Alimento
Definição contextual Espécie de grãos que são produzidos em cachos e cultivados em terrenos pantanosos ou cobertos com água. O arroz.		
Etimologia e/ou Processo de formação *Aba* – cabelo + *ti* – branco *Abá* – homem + *ti* (*ra*) - companheira		
Passagens abonatórias: — A tristeza escurece a vista de Iracema e amarga seu lábio. Mas a alegria há de voltar à alma da esposa, como volta à árvore a verde rama. — Quando teu filho deixar o seio de Iracema, ela morrerá, como o <abati> depois que deu seu fruto. Então o guerreiro branco não terá mais quem o prenda na terra estrangeira. (Iracema) — É tempo de aplacar as iras de Tupã, e calar a voz do trovão. Disse e partiu da cabana. Iracema achegou-se então do mancebo; levava os lábios em riso, os olhos em júbilo: — O coração de Iracema está como o <abati> n'água do rio. Ninguém fará mal ao guerreiro branco na cabana de Araquém. (Iracema)		

Definição dos dicionários consultados:

GD: **Abatí** – s. Milho (p. 389).
TS: n/c
AGC: **abati** *s.m.* Var.: *ubatim, abaty* - Espécie de milho (p. 42).
LCT: **abati** - milho; roça de milho (p. 48).

Nota: José de Alencar traz em nota ao romance *Iracema* (p. 154) que "**abati** é o nome tupi de arroz", entretanto não é o mesmo significado apresentado pelos dicionários consultados. Ambos trazem **abati** como sinônimo de milho, porém Alencar emprega-o como sinônimo de arroz. Embora ambos sejam grãos que servem de alimento, há uma diferença entre eles tanto na forma de cultivo, como nos alimentos oriundos do processamento desses grãos. Arroz e milho não podem ser considerados sinônimos. Outra observação é o fato de que o emprego de **abati** representa dois momentos distintos do sentimento de *Iracema*. Na primeira passagem utiliza **abati** para representar o momento de um fim, ou seja, o **abati** quando lhe são colhidos os grãos, não tem valor e, na água, se esvai e morre; assim como **Iracema** que morrerá após o nascimento do filho. Por outro lado, a segunda passagem representa o momento de alegria de **Iracema** que está em júbilo pela percepção da gravidez, assim como o **abati** diante da possibilidade de produção dos grãos.

Verbete:

Abati. s.m. Alimento. Etim. *Aba* – cabelo + *ti* – branco; *Abá* – homem + *ti* (*ra*) - companheira. Definição contextual: Espécie de grãos que são produzidos em cachos e cultivados em terrenos pantanosos ou cobertos com água. O arroz. Passagens abonatórias: *1) — A tristeza escurece a vista de Iracema e amarga seu lábio. Mas a alegria há de voltar à alma da esposa, como volta à árvore a verde rama. — Quando teu filho deixar o seio de Iracema, ela morrerá, como o <abati> depois que deu seu fruto. Então o guerreiro branco não terá mais quem o prenda na terra estrangeira. (Iracema); 2) — É tempo de aplacar as iras de Tupã, e calar a voz do trovão. Disse e partiu da cabana. Iracema achegou-se então do mancebo; levava os lábios em riso, os olhos em júbilo: — O coração de Iracema está como o <abati> n'água do rio. Ninguém fará mal ao guerreiro branco na cabana de Araquém. (Iracema)*

Ficha N^{o}: 4		
Acajá	s. m.	Elemento da flora
Definição contextual Espécie de fruta da cajazeira, árvore de grande porte que é capaz de formar uma floresta.		
Etimologia e/ou Processo de formação *Acã – iá* – o fruto do caroço cheio.		
Passagens abonatórias: Assim a terra onde nasceu uma floresta de <acajás>, recebe o limo do rio e gera nova floresta mais frondosa que a outra. Jacamim, chama Araci, a filha de nossa velhice. E vós, abarés, chefes, moacaras e guerreiros, segui-me. (Ubirajara)		
Definição dos dicionários consultados: GD: **cajá** – fruto da cajazeira. Riacho de Pernambuco. (p. 527) TS: acayá – s. fruto conhecido, vulgo cajá. (p. 108) AGC: **cajá**. s.m. *Cajá*, *acayá*, *acaiá*, *acaya*, ***acajá***. {T. aka'ia]. Nome comum a diversas plantas da família das anacardiáceas e a seus frutos; *cajazeira*, *cajazeiro*. (p. 85). LCT: **cajá**. fruto da cajazeira, planta da fam. das anacardiáceas. (p. 49)		
Nota: Os dicionários não trazem o vocábulo **acajá** como um verbete, porém AGC o traz como sinônimo de **cajá** e TS emprega o termo **acayá** e explica que é o vulgo **cajá**. Dessa forma, todos os dicionários indígenas trazem uma entrada referindo-se à **cajá**. Alencar não utiliza o vocábulo **cajá**, porém emprega **cajazeira** três vezes no romance *Ubirajara*. O autor, portanto, conhecia a cajazeira, porém preferiu nomear o fruto como **acajá**. A parir das informações não houve a possibilidade de definir com mais precisão.		

Acajá. s.m. Elemento da flora. Etim. *Acã – iá* – o fruto do caroço cheio. Definição contextual: Espécie de fruta da cajazeira, árvore de grande porte que é capaz de formar uma floresta. Passagens abonatórias: *Assim a terra onde nasceu uma floresta de <acajás>, recebe o limo do rio e gera nova floresta mais frondosa que a outra. Jacamim, chama Araci, a filha de nossa velhice. E vós, abarés, chefes, moacaras e guerreiros, segui-me. (Ubirajara)*

Ficha N^{o}: 5		
Acaraú	s. m.	Topônimo
Definição contextual Rio em cujas margens residem os pitiguaras da tribo de Iracema.		
Etimologia e/ou Processo de formação De *acará* - garça + *y* – rio = rio das garças		
Passagens abonatórias: — Volta às margens do <Acaraú>, e teu pé não descanse enquanto não pisar o chão da cabana de Jacaúna. Quando aí estiveres, dize ao grande chefe: — Teu irmão é chegado à taba de seus guerreiros. — E tu não mentirás. O mensageiro partiu. (Iracema) O mancebo cristão viu longe o clarão da festa, e passou além, e olhou o céu azul sem nuvens. A estrela morta, que então brilhava sobre a cúpula da floresta, guiou seu passo firme para as frescas margens do <Acaraú>. (Iracema)		
Definição dos dicionários consultados: GD: n/c TS: **acaraeú** – corr. Acará-hy, pronunciado incorretamente *acará-hú* e *acará-eu*, rio dos acarás; 75, Ceará. Alencar interpretou erroneamente rio das garças confundindo com *agará* a ave vermelha. (, p.107). AGC: n/c LCT: **aracaú** – nome de uma árvore; de **acará-ú**, comida de acará. (p.50)		

Nota:
A definição proposta por TS faz referência a Alencar com a finalidade de esclarecer o erro de explicação. Alencar, em nota ao livro *Iracema*, diz que "**Acaraú** — O nome do rio é *Acaracu* — de *acará* — garça, *co* — buraco, toca, ninho e *y* — som dúbio entre i e u, que os portugueses ora exprimiam de um, ora de outro modo, significando água. Rio do ninho das garças é, pois, a tradução de **Acaracu**; e rio das garças a de **Acaraú**. Usou-se aqui uma liberdade horaciana para evitar em uma obra literária, obra de gosto e artística, um som áspero e ingrato. De resto quem sabe se o nome primitivo não foi realmente Acaraú, que se alterou como tantos outros, pela introdução da consoante?" Alencar serve-se da nota como forma de explicitação da sua pretensão. É possível perceber o tom de réplica do autor ao finalizar a nota com uma interrogação. É possível notar que LCT refere-se a uma árvore, já TS informa que é um rio.

Verbete:

Acaraú. s.m. Topônimo. Etim. *acará* - garça + *y* – rio. Definição contextual: Rio em cujas margens residem os pitiguaras da tribo de Iracema. Passagens abonatórias: *1) — Volta às margens do <Acaraú>, e teu pé não descanse enquanto não pisar o chão da cabana de Jacaúna. Quando aí estiveres, dize ao grande chefe: — Teu irmão é chegado à taba de seus guerreiros. — E tu não mentirás. O mensageiro partiu. (Iracema). 2) O mancebo cristão viu longe o clarão da festa, e passou além, e olhou o céu azul sem nuvens. A estrela morta, que então brilhava sobre a cúpula da floresta, guiou seu passo firme para as frescas margens do <Acaraú>. (Iracema)*

Ficha Nº: 6		
Acauã	s. f.	Elemento da fauna
Definição contextual Ave de rapina da família dos falconídeos que se alimenta de cobras.		
Etimologia e/ou Processo de formação De *caa* – pau, e *uan*, do verbo *u*- comer.		

Passagens abonatórias:
A filha do pajé que ouvira calada, debruçou-se ao ouvido do cristão: — Iracema quer te salvar e a teu irmão; ela tem seu pensamento. O chefe pitiguara é valente e audaz; Irapuã é manhoso e traiçoeiro como a <acauã>. Antes que chegues à floresta, cairás; e teu irmão da outra banda cairá contigo. (Iracema)
Definição dos dicionários consultados: GD: n/c TS: **Acauan**, s. ave conhecida, vulgo, cauan, ave agoreira entre os gentios. (p. 108) AGC: s. m. Var.: *oacaoam*. Cah*uan, macauã, macauhan, macauan, acauán, cauan, cauã, acauan, acauã.* Ave de rapina da família dos falconídeos. (p. 45) LCT: ave grande, também chamada *mucucaguá*, que ataca as serpentes e as matas. Os índios imitavam o seu canto, para afugentar as cobras peçonhentas. Do guarani: *ãcuã*, ligeiro, veloz.
Nota: Não há discrepância entre os dicionários que definem **acauã**, apenas o TD apresenta outra grafia **acauan**. Alencar define **acauã** em nota ao romance Iracema "Ave inimiga das cobras; de *caa* – pau, e *uan*, do verbo *u*- comer" (p. 155) . O autor utiliza **acauã** apenas uma vez em *Iracema*, porém emprega **cauã** mais quatro vezes em seus romances: três vezes em *O Guarani* e uma vez em *Iracema*. Apesar de empregar os dois vocábulos, acauã e cauã, o sentido é mesmo em todos os contextos empregados pelo autor, apesar da decomposição discrepante do significado sugerido por LCT.

Verbete:

Acauã. s.f. Elemento da fauna. Etim. de *caa* – pau, e *uan*, do verbo *u*- comer. Definição contextual: Ave de rapina da família dos falconídeos que se alimenta de cobras. Passagens abonatórias: *A filha do pajé que ouvira calada, debruçou-se ao ouvido do cristão: - Iracema Iracema quer te salvar e a teu irmão; ela tem seu pensamento. O chefe pitiguara é valente e audaz; Irapuã é manhoso e traiçoeiro como a acauã. Antes que chegues à floresta, cairás; e teu irmão da outra banda cairá contigo. (Iracema)*

Ficha N^{o}: 7		
Aimoré	s.m.	Etnônimo
Definição contextual Nome da tribo que atacou a fazenda de D. Antonio de Mariz, pai de Cecília.		
Etimologia e/ou Processo de formação De *hãi*, dentes; *mboré*, preto = os dentes pretos		
Passagens abonatórias: Os gritos dos selvagens responderam de novo: e o canto se prolongou por muito tempo lembrando os feitos gloriosos da nação Aimoré e as ações de valor de seu chefe. (O Guarani) — Guerreiro goitacá, tu és forte e valente; tua nação é temida na guerra. A nação <Aimoré> é forte entre as mais fortes, valente entre as mais valentes. Tu vais morrer. (O Guarani)		
Definição dos dicionários consultados: GD: s. Nome da tribo que habitava o Espírito Santo e Bahia. Nome de rua em S. Paulo. Batista Caetano afirma o significado de dentre pretos (*hãi*, dentes; *mboré*, preto. (p. 514) TS: n/c AGC: n/c LCT: **Aimoré** - etnol. Aguerrida tribo da família linguística Jê, que habitava as margens do Rio Doce (MG) (p. 53)		
Nota: Alencar utiliza **Aimoré** na mesma acepção de GD.		

Verbete**:**

Aimoré. s.m. Etnônimo. Etim. *hãi*, dentes; *mboré*, preto. Definição contextual: Nome da tribo que atacou a fazenda de D. Antonio de Mariz, pai de Cecília. Passagens abonatórias: *1) Os gritos dos selvagens responderam de novo: e o canto se prolongou por muito tempo lembrando os feitos gloriosos da nação Aimoré e as ações de valor de seu chefe. (O Guarani). 2) — Guerreiro goitacá, tu és forte e valente; tua nação é temida na guerra. A nação <Aimoré> é forte entre as mais fortes, valente entre as mais valentes. Tu vais morrer. (O Guarani).*

Ficha N°: 8		
Anajê	Classe gramatical	Elemento da Fauna
Definição contextual Ave de rapina, uma espécie de gavião, cujo voo alcança grandes altitudes.		
Etimologia e/ou Processo de formação *Iana – jê* = o que está separado, o solitário.		
Passagens abonatórias: — Não vale um guerreiro só contra mil guerreiros; valente e forte é o tamanduá, que morde os gatos selvagens por serem muitos e o acabam. Tuas armas só chegam até onde mede a sombra de teu corpo; as armas deles voam alto e direito como o <anajê>. (Iracema) — Vis guerreiros são aqueles que atacam em bando como os caititus. O jaguar, senhor da floresta, e o <anajê>, senhor das nuvens, combatem só o inimigo. (Iracema)		
Definição dos dicionários consultados: GD: n/c TS: **Anagé**, s. o gavião; altera-se em nagé, Bahia (p.109). AGC: **anajê** s.m. Var.: jnagê, enagé, anaje, anagê, anajé, inajé. Ave de rapina (p. 50). LCT: n/c		
Nota: Alencar traz em suas notas do romance *Iracema* "**Anajê** — Gavião". Esta nota seria dispensável se a palavra **anajê** fosse conhecida dos leitores da época, o que nos leva a crer que foi uma criação do autor. Tal afirmação pode ser corroborada com a consulta aos dicionários, os quais apenas TS e AGC, posteriores a Alencar, trazem a palavra entre os verbetes. Como Alencar não explicita o processo de formação, tampouco os dicionários o fazem, leva-nos a conclusão de que o autor criou a palavra e explicitou o significado desejado na nota afirmando que é um gavião.		

Verbete**:**

Anajê. s.m. Elemento da fauna. Etim. *Iana – jê* = o que está separado, o solitário. Definição contextual: Ave de rapina, uma espécie de gavião, cujo voo alcança grandes altitude. Passagens abonatórias: *1) — Não vale um guerreiro só contra mil guerreiros; valente e forte é o tamanduá, que morde os gatos selvagens por serem muitos e o acabam. Tuas armas só chegam até onde mede a sombra de teu corpo; as armas deles voam alto e direito como o <anajê>. (Iracema) . 2) — Vis guerreiros são aqueles que atacam em bando como os caititus. O jaguar, senhor da floresta, e o <anajê>, senhor das nuvens, combatem só o inimigo. (Iracema).*

Ficha Nº: 9		
Andira	s.m.	Antropônimo
Definição contextual Nome do tio de Iracema, grande guerreiro Tabajara.		
Etimologia e/ou Processo de formação *Andi* – espanto, pavor + *rá* – contração de *rahar* – o que faz ou traz		
Passagens abonatórias: O velho <Andira>, irmão do pajé, entrou na cabana; trazia no punho o terrível tacape; e nos olhos uma raiva ainda mais terrível. — O morcego vem te chupar o sangue, se é que tens sangue e não mel nas veias, tu que ameaças em sua cabana o velho pajé. Araquém afastou o irmão: — Paz e silêncio, <Andira>. (Iracema) Ela deve encostar o tacape da luta para tanger o membi da festa. Celebra, Irapuã, a vinda dos emboabas e deixa que cheguem todos aos nossos campos. Então <Andira> te promete o banquete da vitória. (Iracema).		
Definição dos dicionários consultados: GD: **Andirá** – s. Morcego. Cidade do Paraná e prenome de pessoas nada elogioso (p. 515). TS: **andira**, s. o morcego, vampiro (p. 109). AGC: **andirá** s.m. Var.: andura, andira, amdura, andirá. Morcego (p. 52). LCT: **andyrá** – morcego (p. 58).		

Nota:
Para o vocábulo **andira,** há variações em relação à sílaba tônica e todos os dicionários fazem referência ao morcego. GD complementa que, além da cidade do Paraná, é um prenome cuja conotação não é elogiosa. Considerando que GD é um dicionário anterior à publicação de *O Guarani* e que Alencar estudou GD para fazer as anotações, as quais utilizou para criar seus romances, ele nomeia o guerreiro tio de *Iracema* de **andira** por assimilação de comportamentos, ou seja, o morcego **andira** é hematófago por natureza e **Andira**, tio de **Iracema**, saboreia o sangue no sentido conotativo do sentimento de vitória de uma batalha, que provoca, inevitavelmente, morte entre os guerreiros

Verbete:

Andira. s.m. Antropônimo. Etim. Étimo ficcional contextual de Alencar. Definição contextual: Nome do tio de Iracema, grande guerreiro Tabajara. Passagens abonatórias: *1) O velho <Andira>, irmão do pajé, entrou na cabana; trazia no punho o terrível tacape; e nos olhos uma raiva ainda mais terrível. — O morcego vem te chupar o sangue, se é que tens sangue e não mel nas veias, tu que ameaças em sua cabana o velho pajé. Araquém afastou o irmão: — Paz e silêncio, <Andira>. (Iracema). 2) Ela deve encostar o tacape da luta para tanger o membi da festa. Celebra, Irapuã, a vinda dos emboabas e deixa que cheguem todos aos nossos campos. Então <Andira> te promete o banquete da vitória. (Iracema)*

Ao analisarmos as linhas de concordância do vocábulo **andira**, por meio da ferramenta C*oncord* do *WST*, percebemos que é utilizado como antropônimo e como elemento da fauna. Em observação à Figura 65, é possível verificar que, na linha 1, o autor emprega o vocábulo referindo-se ao animal, o que se comprova com a grafia do vocábulo em letra inicial minúscula. Nos outros casos, o autor refere-se ao guerreiro.

Figura 65: Lista de concordância do vocábulo **andira** no *corpus* de estudo

Fonte: A autora, a partir da ferramenta *Concord* do *WST*

Em razão de haver duas acepções para o vocábulo **andira**, há também duas fichas lexicográficas. A duplicidade das fichas se deve ao fato de que o vocabulário a ser disponibilizado *online* trará os vocábulos distribuídos por campos semânticos, ou seja, andira figurará como antropônimo e como elemento da fauna.

Ficha: Nº: 10		
Andira	s.f.	Elemento da Fauna
Definição contextual Animal hematófago que possui hábitos noturnos e se incomoda com a claridade e se alimenta de sangue. É o que hoje se denomina morcego.		
Etimologia e/ou Processo de formação *Andi* – espanto, pavor + *rá* – contração de *rahar* – o que faz ou traz		
Passagens abonatórias: — Conta, virgem das serras, o que sucedeu em teus campos depois que a eles chegou o guerreiro do mar. Iracema referiu como a cólera de Irapuã se havia assanhado contra o estrangeiro, até que a voz de Tupã, chamado pelo pajé, tinha apaziguado seu furor: — A		

raiva de Irapuã é como a <andira>: foge da luz e voa nas trevas. A mão de Poti cerrou súbito os lábios da virgem; sua fala parecia um sopro. (Iracema)
Definição dos dicionários consultados: GD: **Andirá** – s. Morcego. Cidade do Paraná e prenome de pessoas nada elogioso (p. 515). TS: **andira**, s. o morcego, vampiro (p. 109). AGC: **andirá** s.m. Var.: andura, andira, amdura, andirá. Morcego (p. 52). LCT: **andyrá** – morcego (p. 58).
Nota: GD apresenta três acepções para o vocábulo **andira**. Faz referência ao morcego, à cidade e a nome de pessoas, porém não explicita a sua formação. Os demais apresentam definições apenas como morcego. Alencar em nota afirma que "Andira — Morcego; é em alusão a seu nome que Irapuã dirige logo palavras de desprezo ao velho guerreiro" (p. 152), ou seja, neste caso, não se trata de criação do autor.

Verbete:

Andira. s.m. Elemento da fauna. Etim. *Andi* – espanto, pavor + *rá* – contração de *rahar* – o que faz ou traz. Definição contextual: Animal hematófago que possui hábitos noturnos e se incomoda com a claridade e se alimenta de sangue. É o que hoje se denomina morcego. Passagens abonatórias: — *Conta, virgem das serras, o que sucedeu em teus campos depois que a eles chegou o guerreiro do mar. Iracema referiu como a cólera de Irapuã se havia assanhado contra o estrangeiro, até que a voz de Tupã, chamado pelo pajé, tinha apaziguado seu furor: — A raiva de Irapuã é como a <andira>: foge da luz e voa nas trevas. A mão de Poti cerrou súbito os lábios da virgem; sua fala parecia um sopro. (Iracema)*

Ficha N°: 11		
Aracati	s.m.	Diversos Brisa
Definição contextual Nome que os indígenas davam à brisa refrescante vinda do mar.		
Etimologia e/ou Processo de formação De *ara* - vento + *caty* - maresia		

Passagens abonatórias: Era o tempo em que o doce <aracati> chega do mar, e derrama a deliciosa frescura pelo árido sertão. A planta respira; um doce arrepio erriça a verde coma da floresta. (Iracema)
Definição dos dicionários consultados: GD: **Aracati** - Cidade do Ceará. De ara, vento; caty, maresia (p. 517). TS: **Aracaty** – c.orr. ára-caty ou aracatú, vento muito, ou rajada forte; designa, no Ceará, uma cidade das margens do Jaguaribe e um vento mui forte que sopre de repente (p. 111). AGC: **aracati** s.m. Var.: aracaty. Significa este nome bom tempo de ara e catu. Os selvagens do sertão assim chamavam as brisas do mar que sopram regularmente ao cahir da tarde, e correndo pello valla do Jaguaribe se derramam pelo interior e refrigeram da calma abrasadora do sertão. Em nota de José de Alencar. (p. 58-59) LCT: **Aracaty** – brisa do mar, que sopra no Nordeste (Geraldo da Cunha) (p. 64).
Nota: Alencar traz em nota "**aracati** – significa este nome bom tempo – de *ara* e *catu*. Os selvagens do sertão assim chamavam as brisas do mar que sopram regularmente ao cari da tarde e, correndo pelo vale do Jaguaribe, se derramam pelo interior e refrigeram da calma abrasadora do verão" (p. 152). Não há discrepância entre os significados apresentados pelos dicionários e pelo uso de Alencar, tanto na narrativa do romance como em sua nota ao livro Iracema. Note-se que AGC faz referência à definição apresentada por Alencar.

Verbete**:**

Aracati. s.m. Diversos. Brisa. Etim. De *ara* - vento + *caty* – maresia. Definição contextual: Nome que os indígenas davam à brisa refrescante vinda do mar Passagens abonatórias*: Era o tempo em que o doce <aracati> chega do mar, e derrama a deliciosa frescura pelo árido sertão. A planta respira; um doce arrepio erriça a verde coma da floresta. (Iracema)*

Ficha N^{o}: 12		
Araçoia	s.f.	Vestimenta
Definição contextual Espécie de saiote confeccionado com penas coloridas de arara. Era um artefato utilizado por chefes ilustres.		

Etimologia e/ou Processo de formação De arara + *oba* – vestido
Passagens abonatórias: Poti deu a seu irmão o arco e o tacape, que são as armas nobres do guerreiro. Iracema havia tecido para ele o cocar e a <araçóia>, ornatos dos chefes ilustres. (Iracema) Chegou o dia, em que os noivos de Araci deviam disputar a posse da formosa virgem. Era a hora em que o sol transpondo a crista da montanha, estende pelo vale sua <araçóia> d'ouro. A grande nação tocantim cerca a vasta campina. No centro estão os anciões, que formam o grande carbeto. (Ubirajara)
Definição dos dicionários consultados: GD: **Araçóia** - n/c **Araçoyaba** – s. Localidade de São Paulo e de Minas Gerais. De ara, tempo; açoiaba, anteparo contra. Nome dado aos montes isolados, em forma de chapéu. Por extensão, significa o mesmo chapéu que é sempre um anteparo ao tempo. (. 51) TS: **Araçoyá**, contr. araçoyaba. (p. 111) **Araçoyaba**. c. *ara-açoyaba*, cobertura ou anteparo do tempo, e chapéo; ao monte isolado no meuo de uma planície, aos cabeços arredondados dava-se o nome *araçoyaba*; S. Paulo (p. 110-111). AGC: **araçóia** s.f. Var.: *arasaya, arassoia, araçoia*. Espécie de saiote de penas de cores variegadas, usado pelas índias (p. 59). LCT: **Arassóia** – esp. De saiote de penas de cores variegadas, usada pelas índias. (p. 66) SP: n/c
Nota: Alencar explica que araçoia vem de *arara* e *oba* — vestido de penas de arara. Há discrepâncias em relação às definições apresentadas pelos dicionários. GD traz a etimologia por meio de *ara* – tempo + *açoiaba* - anteparo contra. Para ele **araçoyaba** denomina montes em forma de chapéu, assim como o faz TS. Já AGC e LCT propõem a mesma definição que a apresentada pelo Alencar como **araçóia** significando tipo de saiote confeccionado com penas.

Verbete:

Araçoia. s.f. Elemento da flora. Etim. De arara + *oba* – vestido. Definição contextual: Espécie de saiote confeccionado com penas coloridas de arara. Era um artefato utilizado por chefes ilustres. Passagens abonatórias*: 1) Poti deu a seu irmão o arco e o tacape, que são as*

armas nobres do guerreiro. Iracema havia tecido para ele o cocar e a <araçóia>, ornatos dos chefes ilustres. (Iracema). 2) Chegou o dia, em que os noivos de Araci deviam disputar a posse da formosa virgem. Era a hora em que o sol transpondo a crista da montanha, estende pelo vale sua <araçóia> d'ouro. A grande nação tocantim cerca a vasta campina. No centro estão os anciões, que formam o grande carbeto. (Ubirajara)

Ficha Nº: 13		
Araguaia	s.m.	Etnônimo
Definição contextual Nome da nação indígena de Jandira, uma das esposas de Ubirajara.		
Etimologia e/ou Processo de formação De *ara* – arara, papagaio + *guaba* – comida		
Passagens abonatórias: — Ubirajara, o senhor da lança, que empunha o arco da poderosa nação <araguaia>, te manda, a ti, quem quer que sejas, e a todos quantos te obedecem, a sua vontade. (Ubirajara) Araribóia tomou o seu lugar; e o combate prosseguiu com vária fortuna, até Cori que, expelindo o vencedor, manteve-se firme contra todos que vieram disputá-lo. Faltava Jurandir. O estrangeiro avançou gravemente, como convinha a um grande guerreiro da nação <araguaia>. Ele queria dar ao vencedor de tantos combates, o tempo preciso para descansar. (Ubirajara).		
Definição dos dicionários consultados: GD: **Araguaya** – de ara (arara) guaba, comida. Lugar onde os papagaios ou as araras comem. (p. 517) TS: **Araguaya**, c.*araguay*. **Araguay** , c. *araguai*-y, rio do valle dos papagaios. (p.112) AGC: n/c LCT: n/c		
Nota: De acordo com os contextos abonatórios, **araguaia** denomina a tribo indígena de **Jandira**, uma das esposas de **Ubirajara**. Os dois dicionários que trazem o verbete para **araguaia** definem o vocábulo relacionando-o com *papagaios* ou *araras*, porém o processo		

de decomposição de GD e TS para **araguaia** é diferente. Para GD **araguaia** vem de *ara* – arara e *guaba* – comida; já TS diz que a decomposição é *araguai- y* – rio dos papagaios.

Verbete:

Araguaia. s.m. Etnônimo. Etim. De *ara* – arara, papagaio + *guaba* – comida. Definição contextual: Nome da nação indígena de Jandira, uma das esposas de Ubirajara. Passagens abonatórias: *1) — Ubirajara, o senhor da lança, que empunha o arco da poderosa nação <araguaia>, te manda, a ti, quem quer que sejas, e a todos quantos te obedecem, a sua vontade. (Ubirajara). 2) Araribóia tomou o seu lugar; e o combate prosseguiu com vária fortuna, até Cori que, expelindo o vencedor, manteve-se firme contra todos que vieram disputá-lo. Faltava Jurandir. O estrangeiro avançou gravemente, como convinha a um grande guerreiro da nação <araguaia>. Ele queria dar ao vencedor de tantos combates, o tempo preciso para descansar. (Ubirajara).*

Ficha N^{o}: 14		
Boré	s.m.	Utensílio
Definição contextual Instrumento de sopro fabricada a partir do bambu. Espécie de flauta de bambu.		
Etimologia e/ou Processo de formação De corr. *mbyré*, alt. *byré*, *buré*.		
Passagens abonatórias: — Fica tu, escondido entre as igaçabas de vinho, fica, velho morcego, porque temes a luz do dia, e só bebes o sangue da vítima que dorme. Irapuã leva a guerra no punho de seu tacape. O terror que ele inspira voa com o rouco som do <boré>. O potiguara já tremeu ouvindo-o rugir na serra, mais forte que o ribombo do mar. (Iracema) Soaram os <borés>; e ao som do canto de triunfo entoado pelos nhengaçaras, os chefes e os guerreiros saudaram o vencedor dos vencedores. (Ubirajara)		

Os músicos fizeram retroar os <borés>, anunciando o começo da festa; e os servos do amor se estenderam em linha pelo meio da campina. Então os nhengaçaras levantaram o canto nupcial. (Ubirajara)
Definição dos dicionários consultados: GD: n/c TS: **Borê**, ou **buré**, corr. *mbyré*, alt. *byré*, *buré*, o soprado, o que se sopra, gaita do gentio; 122. (p. 116) AGC: **boré** s.m. Espécie de flauta indígena. (p. 74) LCT: **boré** - esp. de flauta indígena (Geraldo da Cunha) (p. 74)
Nota: Alencar traz uma nota para **boré** em *Iracema* "Boré — Frauta de bambu, o mesmo que muré" e não há divergências entre os dicionários consultados.

Verbete**:**

Boré. s.m. Utensílio. Etim. De corr. *mbyré*, alt. *byré*, *buré*. Definição contextual: Instrumento de sopro fabricada a partir do bambu. Espécie de flauta de bambu. Passagens abonatórias: *1) — Fica tu, escondido entre as igaçabas de vinho, fica, velho morcego, porque temes a luz do dia, e só bebes o sangue da vítima que dorme. Irapuã leva a guerra no punho de seu tacape. O terror que ele inspira voa com o rouco som do <boré>. O potiguara já tremeu ouvindo-o rugir na serra, mais forte que o ribombo do mar. (Iracema). 2) Soaram os <borés>; e ao som do canto de triunfo entoado pelos nhengaçaras, os chefes e os guerreiros saudaram o vencedor dos vencedores. (Ubirajara). 3) Os músicos fizeram retroar os <borés>, anunciando o começo da festa; e os servos do amor se estenderam em linha pelo meio da campina. Então os nhengaçaras levantaram o canto nupcial. (Ubirajara).*

Ficha N^{o}: 15		
Itaoca	s.f.	Habitação
Definição contextual Moradia construída de pedras.		
Etimologia e/ou Processo de formação De *itá* – pedra + *oca* - gruta, lapa, casa		

Passagens abonatórias: Os tacapes toparam no ar e os dois guerreiros rodaram como as torrentes impetuosas no remoinho da <Itaoca>. Dez vezes as clavas bateram, e dez vezes volveram para bater de novo. (Ubirajara) A raça dos cabelos do sol cada vez ganhava mais a amizade dos tupinambás: crescia o número dos guerreiros brancos, que já tinham levantado na ilha a grande <itaoca>, para despedir o raio. (Iracema)
Definição dos dicionários consultados: GD: **Itá oca** – parede de pedra. (p. 425) **Itaoca** – s. A caverna, a lapa, a gruta, a casa de pedra. De *itá*, pedra, *oca cova*, gruta, lapa, casa. Rio de Janeiro. (p. 554) TS: **Itaoca**, c. itá— oca, casa de pedra, caverna, furna, lapa. (p. 132) AGC: **itaoca** s.f. Var.: Var. *jtaoca*, *taóca*. Peixe— sapo. (p. 159) LCT: **itaoca** - 1. peixe— sapo. (p. 111) 2. hist. Nome de uma aldeia tupinambá do séc. XVI, nas proximidades da baía da Guanabara. (p. 111)
Nota: Pode-se notar uma certa coincidência entre os dicionários consultados e o sentido utilização por Alencar, como sendo "**Itaóca** — Casa de pedra, fortaleza".

Verbete:

Itaoca. s.f. Habitação. Etim. De *itá* – pedra + *oca* - gruta, lapa, casa. Definição contextual: Moradia construída de pedras. Passagens abonatórias: *1) Os tacapes toparam no ar e os dois guerreiros rodaram como as torrentes impetuosas no remoinho da <Itaoca>. Dez vezes as clavas bateram, e dez vezes volveram para bater de novo. (Ubirajara). 2) A raça dos cabelos do sol cada vez ganhava mais a amizade dos tupinambás: crescia o número dos guerreiros brancos, que já tinham levantado na ilha a grande <itaoca>, para despedir o raio. (Iracema*).

Ficha N^{o}: 16		
Jatobá	s.m.	Antropônimo

Definição contextual

Nome atribuído ao índio líder do grupo dos guerreiros da tribo dos Pitiguaras, pai de Poti e Jacaúna.

Etimologia e/ou Processo de formação

De *jetahy*, *oba* - folha + a – aumentativo = *jetaí* de grande copa
Ieta'i – jataí + *iua* – fruta = o que tem a casca ou superfície dura.

Passagens abonatórias:

— "Antes que o pai de Jacaúna e Poti, o valente guerreiro <Jatobá>, mandasse sobre todos os guerreiros pitiguaras, o grande tacape da nação estava na destra de Batuireté, o maior chefe, pai de <Jatobá>. (Iracema)

<Jatobá> empunhou o tacape dos pitiguaras. Batuireté tomou o bordão de sua velhice e caminhou. Foi atravessando os vastos sertões, até os campos viçosos onde correm as águas que vêm das bandas da noite! (Iracema)

Definição dos dicionários consultados:

GD: **Jatobá** -Árvore. (p. 429)
TS: **Jatobá** - corr.*y- atã* – obá o que tem dura casca, ou a superfície; v. *jatahy*. (p.1360
AGC: **Jatobá** - s.m. var.: *jatubá*. Planta da família das leguminosas, subfamília das cesalpináceas; variedade de jataí. (p. 176)
LCT: v. **jetaýba**. (p. 177)
Jetaýba – jataí, jatobá; grande árvore da fam. das leguminosas. (p. 119)

Nota: Alencar traz duas explicações para **jatobá** em suas notas ao romance *Iracema*: 1) "Jatobá — Grande árvore real. O lugar da cena é o sítio da hoje Vila Viçosa, onde diz a tradição ter nascido Camarão" (p. 156) e 2) "**Jatobá** — Árvore frondosa, talvez de *jetahy*, *oba* — folha, e a, aumentativo; *jetaí* de grande copa. É o nome de um rio e de uma serra em Santa Quitéria" (p. 157). Merece destaque o que TS apresenta o processo de formação da palavra, "*y- atã – obá* o que tem dura casca", sem mencionar à árvore. Ressaltamos que nomeia a árvore e o fruto dessa árvore *jatobá*. O autor faz associação entre a realeza da árvore com o fato de o **Jatobá**, personagem do romance, ser também líder na tribo. Não é

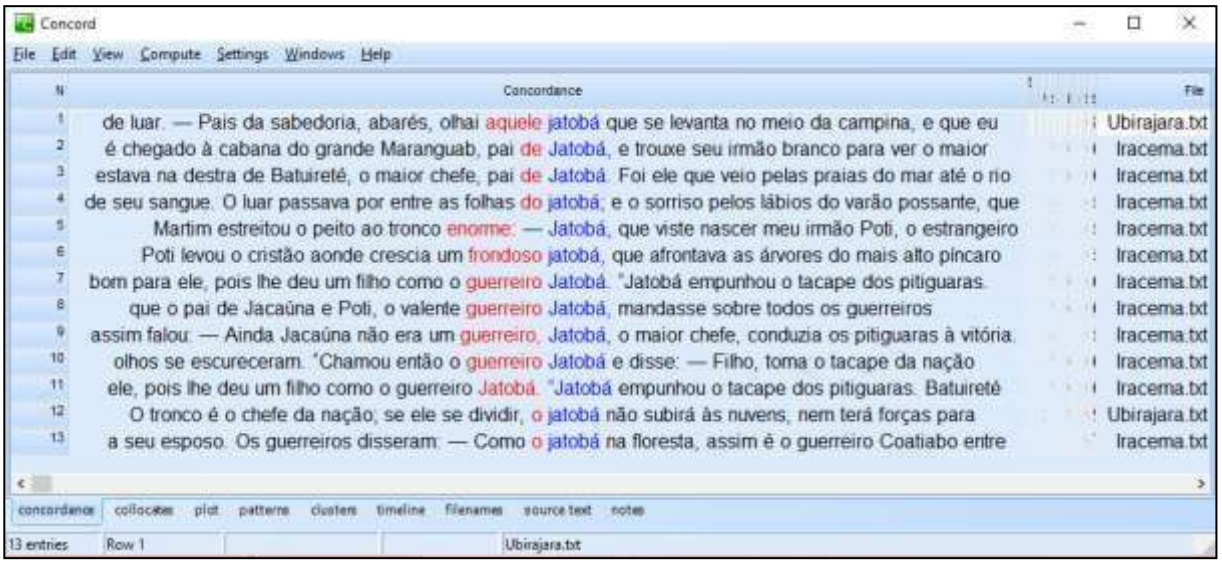
possível afirmar se o autor associa com o sentido apresentado por TS de casca dura, já que ele não menciona esse dado em sua nota

Verbete:

Jatobá. s.m. Elemento da flora. Etim. Étimo ficcional contextual de Alencar. Definição contextual: Nome atribuído ao índio líder do grupo dos guerreiros da tribo dos Pitiguaras, pai de Poti e Jacaúna. Passagens abonatórias: *1) — "Antes que o pai de Jacaúna e Poti, o valente guerreiro <Jatobá>, mandasse sobre todos os guerreiros pitiguaras, o grande tacape da nação estava na destra de Batuireté, o maior chefe, pai de <Jatobá>. (Iracema). 2) <Jatobá> empunhou o tacape dos pitiguaras. Batuireté tomou o bordão de sua velhice e caminhou. Foi atravessando os vastos sertões, até os campos viçosos onde correm as águas que vêm das bandas da noite! (Iracema)*.

Alencar utiliza o vocábulo **jatobá** em duas acepções: como elemento da flora e como antropônimo. A seguir apresentamos a Figura 66 com a qual demonstramos as linhas de concordâncias do vocábulo **jatobá**. Como é evidente pelas linhas 2, 3, 5, 7. 8, 9, e 11 que o autor nomeia o guerreiro da tribo com esse nome, o que é evidente, além do contexto abonatório a utilização da letra maiúsculo no início do vocábulo. Porém, nas linhas 1, 4, 6, 12 e 13, é notória a referência à árvore. Ressaltamos que o autor não faz referência ao fruto também de mesmo nome.

Figura 66: Linhas de concordância do vocábulo **jatobá**



Fonte: A autora, a partir da ferramenta C*oncord* do *WST*

Ficha Nº: 17		
Jatobá	s.m.	Elemento da Flora

Definição contextual:

Árvore da família das leguminosas que possui o fruto com uma casca dura no formato de uma grande vagem que se assemelha a um estojo, capaz de cair de grande altitudes e não se quebrar. Dentro do fruto, a polpa tem o formato semelhante a uma banana. O fruto da árvore também leva o nome de jatobá.

Etimologia e/ou Processo de formação

De *jetahy*, *oba* - folha + a – aumentativo = *jetaí* de grande copa
Ieta'i – jataí + *iua* – fruta = o que tem a casca ou superfície dura.

Passagens abonatórias:

Poti levou o cristão aonde crescia um frondoso <jatobá>, que afrontava as árvores do mais alto píncaro da serrania, e quando batido pela rajada, parecia varrer o céu com a imensa copa. (Iracema)

— Pais da sabedoria, abarés, olhai aquele <jatobá> que se levanta no meio da campina, e que eu só posso ver agora na sombra de minha alma. (Ubirajara)

As grossas raízes são os abarés que sustentam o chefe com o seu conselho. Os galhos fortes são os moacaras que cercam o chefe e geram a multidão de guerreiros mais numerosa que as folhas das árvores. O tronco é o chefe da nação; se ele se dividir, o <jatobá> não subirá às nuvens, nem terá forças para resistir ao tufão. (Ubirajara).

Definição dos dicionários consultados:

GD: **Jatobá** -Árvore. (p. 429)
TS: **Jatobá** - corr.*y*- *atã* – obá o que tem dura casca, ou a superfície; v. *jatahy*. (p.1360
AGC: **Jatobá** - s.m. var.: *jatubá*. Planta da família das leguminosas, subfamília das cesalpináceas; variedade de jataí. (p. 176)
LCT: v. **jetaýba**. (p. 177)
Jetaýba – jataí, jatobá; grande árvore da fam. das leguminosas. (p. 119)

Nota: Todos os dicionários remetem ao elemento da flora, porém apenas o TS refere-se ao fruto da árvore.

Verbete:

Jatobá. s.m. Elemento da flora. Etim. De *jetahy*, *oba* - folha + a – aumentativo = *jetaí* de grande copa. Definição contextual: Árvore da família das leguminosas que possui o fruto com uma casca dura no formato de uma grande vagem que se assemelha a um estojo, capaz de cair de grande altitudes e não se quebrar. Dentro do fruto, a polpa tem o formato semelhante a uma banana. O fruto da árvore também leva o nome de jatobá. Passagens abonatórias: *1) Poti levou o cristão aonde crescia um frondoso <jatobá>, que afrontava as árvores do mais alto píncaro da serrania, e quando batido pela rajada, parecia varrer o céu com a imensa copa. (Iracema). 2) — Pais da sabedoria, abarés, olhai aquele <jatobá> que se levanta no meio da campina, e que eu só posso ver agora na sombra de minha alma. (Ubirajara). 3) As grossas raízes são os abarés que sustentam o chefe com o seu conselho. Os galhos fortes são os moacaras que cercam o chefe e geram a multidão de guerreiros mais numerosa que as folhas das árvores. O tronco é o chefe da nação; se ele se dividir, o <jatobá> não subirá às nuvens, nem terá forças para resistir ao tufão. (Ubirajara).*

Importante ressaltar que a definição de elemento da flora e da fauna baseia-se em características distintivas de outros elementos ou em semelhanças. As definições, portanto, foram baseadas no conhecimento enciclopédico de Alencar, já que as estruturas da fauna e da flora não refletem estruturas linguísticas, mas representam características do mundo real, razão pela qual "os dicionários unilíngues têm grandes dificuldades para definir linguisticamente os vocábulos em questão e devem recorrer, para tanto, à terminologia científica ou a descrições e imagens dos objetos designados". (COSERIU, 1977 apud ABBADE, 2011).

Já os antropônimos e topônimos, como mencionado, a definição se baseia nas características dos personagens, conforme afirmam Eckert e Röhrig (2016, p. 176) que, no caso de Alencar, um nome "dificilmente é usado como mera etiqueta de identificação, tal como ocorre com a maioria dos nomes atribuídos aos brasileiros na atualidade". Pode-se atestar o dito, por exemplo, em *Iracema,* quando **Martim** muda o nome para **Coatiabo**, ao ser

aceito pela tribo de **Iracema**. Igualmente ocorre com o personagem do romance *Ubirajara.* Eckert e Röhrig (2016) analisam os antropônimos de *Ubirajara* a afirmam que os nomes são motivados pelas características dos personagens como ocorre com **Jaguarê**, que significa aquele "que tinha vencido todos os animais, até o temido jaguar", conforme Alencar explica no próprio romance. Segundo os autores, era desonroso entre os indígenas perguntar o nome de um estrangeiro, assim eles, com base nas características, escolhiam um nome. Não ocorreu com **Jaguarê**, que por sua própria vontade escolhe seu novo nome com a frase: "– Eu sou aquele que veio trazido pela luz do céu. Chama-me **Jurandir**". Alencar então tem o cuidado de explicar a origem no nome Jurandir: formado a partir da contração da frase "*Ajur-rendi-pira*, o que veio trazido pela luz". Do tupi *jurandira*, de *jura* (boca) e *ndiera* (abelha de mel), significa: boca melíflua. Que pode ser entendido como aquele de palavras doces. Ao derrotar **Pojucã** utilizando sua lança, Jaguarê proclama seu novo nome, conforme Alencar declara no corpo do seu romance: "Eu sou **Ubirajara**, o senhor da lança, o guerreiro terrível que tem por arma uma serpente". (ECKERT; RÖHRIG, 2016, p. 182-183).

Assim, tratamos das definições baseadas nas obras literárias de Alencar, considerando, principalmente, os contextos de usos no corpo dos textos. Após a apresentação das fichas lexicográficas e os verbetes que servem de modelo para a elaboração das demais fichas e dos demais verbetes, que darão origem ao vocabulário a ser disponibilizado online, passamos para a proposta de elaboração de um vocabulário *online*.

7.3 Proposta para elaboração de Vocabulário *online* do léxico indianista de José de Alencar

A elaboração de obras lexicográficas exige do profissional, imbuído dessa tarefa, um planejamento minucioso do seu trabalho, seguido do cumprimento das etapas que sucedem à execução das atividades que culminam na obra. Quando se propõe à elaboração de um dicionário, de um glossário ou de um vocabulário eletrônicos ou para fins de disponibilização *online,* o apoio da informática é essencial, evidentemente, pela especificidade. Assim sendo, a união entre o trabalho lexicográfico e a informática poderá ser frutuosa. O que é impedido pelo formato em papel, dada a rigidez de organização e disposição dos verbetes, é possibilitado pelo formato *online*, como a exploração dos textos lexicográficos por meio das possibilidades de se anexar informações por meio de *hiperlinks*.

Quando se trata de um dicionário ou vocabulário disponível *online*, as formas de consulta e os verbetes podem ser constantemente alargados pelo lexicógrafo, além de possibilitar a exploração de acordo com as necessidades do consulente, por exemplo, a identificação do campo semântico, da etimologia, de contextos abonatórios, enfim o que se dispuser no ato da elaboração.

É conveniente que os dicionários eletrônicos possuam características específicas, porém eles ainda esbarram nas mesmas características dos impressos, pois como afirma Van Campenhoult (2000)

> os responsáveis pela difusão dos dados já existentes nos dicionários impresso em papel se acharam na necessidade de convertê-los de maneira que possam ser explorados a partir da interface de interrogação. De fato, a grande maioria dos dicionários disponíveis no mercado ainda é redigida em função do modo de apresentação herdado da tradição lexicográfica e destinado a facilitar a consulta sobre um suporte escrito. (VAN CAMPENHOULT, 2000 apud SILVA, 2008, p.89).

Os dicionários alfabéticos impressos já são familiares aos consulentes, pois esta foi a primeira estrutura com a qual eles tiveram os primeiros contatos e isso permaneceu por muitos anos, até surgirem os dicionários eletrônicos. Não havia interesse em publicar dicionários especializados impressos por considerar que, para esse produto, havia um público-alvo reduzido, que são especialistas da área, por isso há um número reduzido desse tipo de material.

Em razão dessa raridade, os consulentes de áreas específicas deveriam recorrer aos dicionários de termos gerais da língua os quais, em grande parte, não dão conta de definir as especificidades das palavras. Por exemplo, como lembra Barbosa (2004), um vocábulo como ‘navegar’ pode significar “viajar pela água, com embarcação” que difere da utilização nos domínios da aeronáutica e da informática.

Pela mesma razão da especificidade, também não há interesse em comercializar os vocabulários na forma impressa. Daí que a disponibilização dos vocabulários *online* se torna um grande favorecimento para esse mesmo público que necessita de informações mais pontuais sobre algum vocábulo. Isso porque, quanto mais comercial um dicionário, mais investimentos são alocados para construção desse modelo.

Os dicionários eletrônicos possuem características diferenciada dos impressos. Não há o posicionamento dos verbetes em ordem alfabética, porque o consulente digita a palavra desejada no campo de busca e poderá obter os campos de informações como etimologia, contexto abonatório, campo semântico, dentre outros. O suporte eletrônico propicia a busca e

a manipulação das informações com mais agilidade, além de ser possível ao consulente escolher, dentre os campos de informações disponíveis, o que se deseja obter de informações, por exemplo, etimologia da palavra, sinônimos, antônimos, dentre vários.

A disponibilização de um vocabulário em meio *online* requer um tratamento especializado para transposição dos dados, que antes eram disponibilizados em meio impresso, para o meio eletrônico. Essa disponibilização não pode ser apenas uma transposição de dados, mas requer uma análise dos recursos que o meio eletrônico-computacional oferece, incorporando a dinamicidade que os recursos tecnológicos possibilitam.

Ressaltamos que se trata de uma proposta, que será desenvolvida em momento posterior à finalização desta tese. Por meio da nossa pesquisa, já obtivemos a lista dos vocábulos indianistas que comporão o vocabulário *online*, porém nosso objetivo para esta tese é apresentar uma proposta, em razão disso, o modelo apresentado do vocabulário poderá sofrer alterações para a versão final, quando obtivermos as contribuições dos profissionais da informática que trabalharão nos procedimentos de produção e disponibilização dos dados *online*.

Esta seção é dedicada ao direcionamento das etapas de elaboração de nossa proposta, assim como se propõe a delinear as possíveis formas de consulta do vocabulário *online*. Reforçamos a ideia de que os dicionários e vocabulário especializados requerem uma metodologia adequada às descrições dos vocábulos. A seguir, apresentamos a arquitetura da proposta por nós delineada.

Na página inicial, mostraremos o título do Vocabulário com os ícones *Início*, *Apresentação*, *Campos Semânticos*, *Palavras*, o campo de busca e as letras do alfabeto. Embora tenhamos desenvolvido nosso trabalho utilizando o conceito de vocábulo que o distingue de palavra, optamos por manter no vocabulário *online* **Palavras+**, por considerar que, para o consulente, seria mais familiar. O consulente terá a opção de consulta por três meios: escolha por campo semântico, por palavras, pela escolha das letras do alfabeto que possuem pelo menos uma palavra, ou pela digitação de uma palavra já conhecida por ele. As letras alfabéticas que estão destacadas em cor azul não possuem vocábulos indígenas que iniciam por elas, assim sendo, não há nenhum vocábulo indígena iniciado por D, F, K, L, R, V, W e Y. A Figura 67 apresenta a página inicial do vocabulário.

Figura 67: Vista da página inicial do Vocabulário *online*

Fonte: A autora

Ao escolher a opção "apresentação", o consulente obterá informações sobre a justificativa do vocabulário, sobre o autor José de Alencar, sobre a tríade indianista ou visualizar as referências utilizadas para elaboração do vocabulário. Já o menu Ca*mpos Semânticos*, destina-se a visualizar a lista dos campos semânticos identificados: Flora, Fauna, Topônimos, Antropônimos, Etnônimos, Utensílios/vestimenta, Alimentos, Espiritualidade, Habitação e Diversos. O consulente poderá, também, optar pela lista em ordem alfabética dos vocábulos, escolhendo a opção *Palavras*. Neste caso, todos os vocábulos indianistas poderão ser visualizados rolando a caixa disponibilizada na tela. Ressaltamos que as opções não ocorrem simultaneamente, ou seja, o consulente deverá escolher entre visualizar os dados da apresentação, dos campos semânticos ou os vocábulos. A Figura 68 apresenta as possibilidades descritas.

Figura 68: Vista do Vocabulário a partir da escolha de ícones

Fonte: A autora

Pretendemos que a página possibilite a navegação do usuário de forma dinâmica e dê livre escolha à forma de acesso aos vocábulos, dentre as possibilidades arquitetadas.

Avançando as possibilidades de pesquisa, o consulente poderá optar pelo campo semântico em que estarão listadas todas as palavras pertencentes a cada um para, a partir da escolha visualizar a definição. Conforme salientado por Dubois (2007, p. 89), definição "é a análise semântica da palavra de entrada. Consta de uma séria de paráfrases sinonímicas da palavra de entrada (...) a definição é uma descrição desse objeto, tal como este é secionado no mundo pelo léxico de uma língua".

No nosso vocabulário, as definições são contextuais com base nos contextos abonatórios abstraídos dos romances indianistas de Alencar: *Iracema*, O *Guarani* e *Ubirajara*. Essas definições buscam fornecer as principais características dos vocábulos no contexto das narrativas. Considerando os contextos de usos dos vocábulos indianistas, propusemos uma ordenação por campos semânticos, conforme apresentado no Quadro 11 em que se relacionam os 367 vocábulos indianistas identificados.

Quadro 11: Distribuição dos vocábulos indígenas por campo semântico

Antropônimos	**Etnônimos**	**Topônimos**	**Flora**	**Fauna**
- Abaeté	- Aimoré	- Acaraú	- Abacaxi	- Acauã
- Aimoré	- Araguaia	- Acarú	- Acajás	- Anajê
- Andira	- Caraíba	- Apodi	- Açaí	- Andira
- Aquiraz	- Caramuru	- Aratanha	- Acaris	- Ará
- Araci	- Emboabas	- Aratuba	- Aguapés	- Arara
- Araquém	- Goitacá	- Arroio	- Aipim	- Arirama
- Ararê	- Guaraciaba	- Campina	- Airi	- Ariranha
- Araribóia	- Guarani	- Camucim	- Ananás	- Araruna
- Aresqui	- Marabá	- Capoeira	- Andiroba	- Ati
- Batuirité	- Moacara	- Ceará	- Angico	- Atiati
- Boitatá	- Morubixaba	- Guaiúba	- Anum	- Boicininga
- Cacique	- Nhengaçara	- Ibiapaba	- Araçá	- Caipora
- Camacã	- Nhengaíba	- Ibiapina	- Araribás	-Caititu
- Camoropim	- Pitiguara	- Icó	- Aroeira	- Caitetu
- Canicrã	- Potiguara	- Iguape	- Bambu	- Camoropim
- Caraúba	- Tabajara	- Ipu	- Banana	- Camorupim
- Cauatá	- Tamoios	- Jacarecanga	- Biribá	- Campeiro
- Caubi	- Tapuia	- Jacobina	- Cabuíba	- Caninana
- Cearense	- Tapuitinga	- Jaguaribe	- Cajazeira	- Capivara
- Ceci	- Tocantim	- Jereraú	- Caju	- Carcará
- Coatiabo	- Tupi	- Mairi	- Cajueiro	- Cauã
- Crebã	- Tupinambá	- Maranguab	- Carnaúba	- Coati
- Curumim		- Maranguape	- Capim	- Coruja
- Guaribu		- Mata	- Catuíba	- Crajuá
- Iara		- Mearim	- Cipó	- Crauatá
- Iracema		- Meruoca	- Copaíba	- Crautá
- Irapuã		- Mocejana	- Crajurú	- Craviri
- Itaquê		- Mocoripe	- Craúba	- Croás
- Itaquêe		- Mundaú	- Embaíba	- Cuandu
- Jacamim		- Muritiapuá	- Guabiroba	- Cumari
- Jacaúna		- Pacatuba	- Guaraná	- Cupim
- Jaci		- Pacoti	- Guaximas	- Cutia
- Jaguaraçu		- Pará	- Igapé	- Gará
- Jaguarê		- Paraíba	- Imbé	- Goaná
- Jandira		- Parnaíba	- Imbu	- Goiamum
- Jatobá		- Pirapora	- Jacarandá	- Graúna
- Javari		- Porongaba	- Jataí	- Guará
- Jurandir		- Potengi	- Jatobá	- Guanumbi
- Jurupari		- Quixeramobim	- Jenipapo	- Guaxinim
- Jutaí		- Sapiranga	- Jequeriti	- Inhuma
- Jutorib		- Tanatinga	- Jetaí	- Intanha
- Maranguab		- Tauape	- Jetica	- Irara
- Moacir		- Trairi	- Juazeiros	- Jaburu
- Mocoribe		- Uruburetama	- Juçara	- Jabuti
- Ogib		- Soipé	- Juquiri	- Jaçan

- Pahã - Pajé - Paquequer - Peri - Pirajá - Pojucã - Poti - Saí - Tubim - Ubirajara - Uiraçu		- Taba - Xingu	- Jurema - Manacá - Mandioca - Mangaba - Mangabeira - Maniva - Maracujá - Monguba - Muriti - Mururê - Oiticica - Ouricuri - Pequiá - Piaçaba - Pirijá - Pitanga - Sapé - Sapucaia - Taioba - Taquari - Ubaia - Ubiratã - Urutaí - Uvaia	- Jaçanã - Jacu - Jaguar - Jandaia - Japi - Japim - Jararaca - Jati - Jiboia - Juriti - Juruti - Majoí - Maracajá - Mutuca - Nambu - Nandu - Oitibó - Paca - Papagaio - Piau - Piranha - Poraquê - Quati - Sabiá - Sagui - Saí - Saixê - Sapoti - Saúva - Sucuri - Tamanduá - Tapir - Tatu - Teiú - Traíra - Tucano - Tuim - Uraçá - Urubu - Urutau - Zabelê
Utensílios/ vestimentas	**Alimentos**	**Espiritualidade**	**Habitação**	**Diversos**
- Araçóia - Balaio - Boré - Caiçara - Camuci - Camucim	- Abati - Biaribi - Carimã - Cauim - Jurema - Mingau	- Abaré - Anhanga - Jaci - Jurupari - Tupã	- Choça - Itaoca - Oca - Ocara - Taba	- Aracati - Corisco - Guarani - Irerê - Maracatim - Maranduba

- Carioba - Clavina - Cumbuca - Igaçaba - Igara - Inúbia - Jequis - Jirau - Maracá - Maracatim - Membi - Muçurana - Murinhém - Piroga - Sapopema - Taboca - Tacape - Tamandaré - Tangapema - Taquara - Trocano - Uiraçaba - Uru	- Moquém - Piracém			- Pocema - Pororoca - Sopé - Ticum

Fonte: A autora

Ressaltamos que não se descarta a possibilidade de que alguns vocábulos pertençam a mais de um campo semântico, uma vez que, para os efeitos da presente tese, não foram elaboradas as fichas lexicográficas da totalidade dos 367 vocábulos.

Um possível passo para a consulta é o consulente escolher a busca clicando sobre o menu campo semântico Flora, por exemplo, que listará todos os vocábulos pertencentes àquele campo semântico. A Figura 69 demonstra um exemplo de uma escolha pelo campo semântico Flora. Assim, o consulente poderá escolher entre os vocábulos listados aquele que deseja consultar as informações disponíveis.

Figura 69: Vista dos campos semânticos

Fonte: A autora

Valendo-se dos artifícios computacionais, propomos também uma busca pela palavra, por exemplo, **jatobá**. Se a escolha for esta possibilidade, o consulente obterá as informações sobre a palavra, conforme exemplificado na Figura 70.

Figura 70: Dados sobre o vocábulo **jatobá** a partir da escolha do consulente

Fonte: A autora

A pesquisa poderá ser realizada também digitando a palavra desejada no campo de busca que abrirá a tela com as informações desejadas, conforme os dados do vocábulo **jatobá** na Figura 36. Assim, poderá optar pelo menu *Palavras* e, ao tomar esse procedimento, o consulente visualizará todos os vocábulos e poderá escolher o que deseja e clicar sobre ele. Fazendo isso, abrirá a tela como exemplificado na Figura 71 em relação ao vocábulo **abacaxi**.

Figura 71: Vista da tela do computador a partir da consulta pelas palavras

Fonte: A autora

Ressaltamos que, por se tratar de uma proposta de vocabulário que será disponibilizada *online*, não tratamos de questões relacionadas aos direitos autorais. Portanto, posteriormente, todos os procedimentos necessários serão tomados, como elaboração de um *layout*, escolha e cores e imagens e os trâmites legais de hospedagem.

Entendemos que, por se tratar de uma proposta, o vocabulário poderá sofrer adaptações para buscar o equilíbrio entre as necessidades dos consulentes e as informações que serão necessárias para satisfazê-los. Portanto não consideramos um produto acabado uma vez que os procedimentos de elaboração vão ao encontro dos interesses dos usuários, o que somente será possível com a continuação deste trabalho. Deverá, por fim, atentar para as necessidades reais dos usuários e propor informações que sejam o mais próximo possível de satisfazer a essas necessidades.

Outro aspecto relevante é que a sociedade está em constante mudança em termos de aquisição de informação, principalmente, o que foi motivado pelo avanço das tecnologias digitais. Por exemplo, é muito comum os leitores substituírem os livros impressos pelos digitais, cuja leitura é realizada em qualquer ambiente em seus equipamentos eletrônicos manuais como celulares e *tablets*, sendo assim os dicionários e vocabulários também devem

acompanhar essa nova realidade. Não há sentido em um leitor ler o livro em formato digital e pesquisar uma palavra em um dicionário impresso, por exemplo. Assim sendo, os dicionários e vocabulários deste século deverão refletir a dinamicidade e a flexibilidade em razão das mudanças das últimas décadas.

No caso de nossa proposta, trata-se de uma fatia peculiar do léxico, ou seja, especificamente o léxico indianista em José de Alencar. O vocabulário, apesar de constar de vocábulos específicos das obras de Alencar, poderá trazer informações que extrapolam os limites dos romances, tendo em vista que os vocábulos indígenas fazem parte do léxico geral da Língua Portuguesa. Muitos deles saíram das obras de Alencar para ganharem notoriedade entre os falantes da língua portuguesa, como é o caso de **Iracema**, **Moacir**, alguns topônimos, nomes relacionados à fauna, como **abacaxi**, **jatobá**, **aroeira**, **banana** e outros como **coruja**, **choça** e **balaio**.

Finalizamos este capítulo, em que apresentamos 17 fichas lexicográficas de vocábulos indianistas com seus respectivos verbetes, e no qual esboçamos uma proposta de um vocabulário indianista de Alencar, com a finalidade de disponibilização *online*. A seguir, traçaremos as conclusões e algumas considerações sobre nossa pesquisa, como também revisitamos os objetivos da tese e propomos algumas questões para futuros estudos. Por fim, levantamos questões relacionadas às limitações deste trabalho.

CONCLUSÕES E CONSIDERAÇÕES

Ao longo da trajetória desta pesquisa, o percurso investigativo sofreu algumas alterações de foco, pois, em princípio, nosso propósito era elaborar um Vocabulário com o léxico indianista de Alencar para disponibilização *online*. Porém, com o manuseio do *corpus*, com as orientações da banca de qualificação e do orientador da pesquisa, e com o aprofundamento teórico, a análise do léxico pelo viés da Etimologia configurou-se, também, um campo profícuo de pesquisa. Sendo assim, desdobramos nossa pesquisa em duas perspectivas: (i) identificação do léxico indianista nas obras também consideradas indianistas de Alencar, por meio do contraste com as outras obras do mesmo autor; (ii) realização de uma análise etimológica ficcional contextual desse léxico indianista. Para além desses desdobramentos, (iii) a proposição de um desenho para elaboração de um Vocabulário com os vocábulos indígenas do autor, para disponibilização *online*.

Revisitamos, a seguir, os objetivos de pesquisa da tese, a fim de destacar os pontos que se afiguram como relevantes para as conclusões e considerações de nossa pesquisa.

Como primeiro objetivo, propusemo-nos a:

i) Identificar e descrever o léxico específico de Alencar em suas obras indianistas.

Consideramos este objetivo alcançado com base na apresentação dos dados no Capítulo 5 "Análise contrastiva do *corpus* de estudo de Alencar com os *corpora* de referência", quando apresentamos as convergências entre as listas de palavras-chave entre o *corpus* de estudo e os demais *corpora* de referência.

Alencar utiliza um léxico específico em suas obras indianistas, se considerado que as palavras-chave de maior chavicidade do *corpus* de estudo, em relação aos quatro *corpora* de referência, são praticamente as mesmas. Esta regularidade se manteve, apesar de os *corpora* de referências apresentarem características diversas quanto à extensão. Por exemplo, o CorpRef-AcadTeses possui 620.068 formas e o CorpRef-Alencar com 61.121, diferença que não alterou, significativamente, a lista de palavras-chave. Outro aspecto, em relação à composição dos *corpora* de referência, foi a formação em relação aos gêneros dos textos compilados. Isso porque, por exemplo, o CorpRef-Alencar é formado pelos textos do próprio Alencar, predominantemente por obras literárias e o CorpRef-AcadTeses formado por um

banco de teses. Entretanto, essa diferença não afetou as listas de palavras-chave, cujas palavras consideradas chave se mantiveram as mesmas, em porcentagem significativa.

A despeito da diversidade dos *corpora* de referência, as palavras-chave do *corpus* de estudo mantiveram a regularidade nos diferentes contrastes estabelecidos, corroborando a premissa de que Alencar utiliza um léxico específico em suas obras indianistas.

Além disso, identificamos uma alta produtividade, no aspecto criação e emprego do léxico indianista, por parte de Alencar, em suas obras indianistas. O primeiro aspecto dessa produtividade refere-se à quantidade de vocábulos indianistas identificados no *corpus* de estudo, ou seja, um total de 367 vocábulos. Ressaltamos que as ferramentas do *WST* nos mostraram que o *corpus* de estudo possui um total de 758 vocábulos substantivos e adjetivos, dentre os quais 367 são indianistas, o que corresponde a, aproximadamente, 41%, isto é, quase a metade dos substantivos e adjetivos utilizados por Alencar em *Iracema*, *O Guarani* e *Ubirajara* são vocábulos de origem indígenas ou étimos indígenas do autor, conferindo o caráter de produtividade e ineditismo lexical em suas obras.

Soma-se à pujança dos vocábulos indígenas empregados pelo autor, a quantidade de vocábulos com frequência um ou zero no *corpus* de estudo em relação aos *corpora* de referência. Como exemplo, ao se contrastar o *corpus* de estudo com o CorpRef-Alencar, verificamos que dos 291 itens identificados por este contraste, 55 possuem a frequência zero no *corpus* de referência, já no contraste com o CorpRef-Nov, dos 477 itens, 104 possuem a frequência zero. Esses dados são expressivos para justificar que Alencar não dispensou esforços para utilizar um léxico específico em suas obras indianistas.

Ainda um fator relevante em relação à produtividade do autor, foi o quantitativo de 57 vocábulos que não constam dos dicionários consultados, porém foram considerados indígenas por nomear elementos relacionados ao universo dos índios. Novamente, o processo criativo de Alencar se sobressaiu para atender aos seus desejos literários.

Para o tratamento adequado dos dados lexicográficos adequadamente, inicialmente, compilamos os textos que compuseram o *corpus* de estudo, as três obras indianistas de Alencar: O *Guarani*, *Iracema* e *Ubirajara*, e o *corpus* de referência formado com as demais obras do mesmo autor. Continuamos os procedimentos, promovendo a limpeza do *corpus*, a fim de corrigir as imperfeições dos textos e remover elementos que não comporiam os *corpora*, dentre eles, os caracteres não inteligíveis.

O Programa *WordSmith Tools*, versão 6.0, revelou ser essencial em nossa investigação, no sentido de identificar se José de Alencar utiliza um léxico específico e indianista em suas obras consideradas indianistas. Essa investigação se veria impossibilitada, se o trabalho fosse realizado sem o auxílio de um programa computacional potente, para auxiliar nas análises lexicais.

O segundo objetivo a que nos propusemos foi

ii) Realizar uma análise etimológica ficcional contextual com base nos itens lexicais indígenas de Alencar.

Como pode ser observado no Capítulo 6 "A Etimologia Ficcional Contextual: a criação de vocábulos indígenas", alcançamos o objetivo proposto. Inicialmente, com base nas teorias sobre Etimologia, nos conceitos sobre ficção e contexto, elaboramos a expressão "Etimologia Ficcional Contextual", para dar conta da análise das criações dos étimos por Alencar, com base em seus propósitos contextuais. Consideramos Etimologia Ficcional Contextual, porque Alencar cria os étimos para imprimir aos romances aspectos do seu intuito, no que se refere à criação literária e à emancipação da língua portuguesa brasileira, em relação aos europeus.

Como procedimento para criação de seus étimos, Alencar, inicialmente, com base em estudos sobre os vocábulos indígenas, decompunha os vocábulos indígenas que conhecia por meio de estudos, como, por exemplo, o vocábulo **igara**. Esse vocábulo é citado por Aires de Casal, uma das possibilidades de contato do autor com o vocábulo. Alencar, então, com base no dicionário de GD, que traz os fragmentos: *ig* – água e *jara* – senhor, decompõe o vocábulo **igara** em *ig* + *jara* = senhor ou senhora das águas. Em seguida, ele cria outros vocábulos por meio da junção de fragmentos e lhes atribui significados contextuais no interior da obra.

Esse foi o procedimento adotado por Alencar de atribuir aos vocábulos criados, o significado apropriado de acordo com o romance, ou seja, se o autor desejava atribuir um significado a um personagem, ele inventava um nome e o explicava, considerando as características desse personagem nos romances. Assim também o fez em relação aos diversos vocábulos empregados, que compõem os outros campos semânticos. Alencar, segundo Schwamborn (1998), afirmou que os apontamentos do autor serviram como material para suas próprias etimologias e formação de palavras, que contribuíram para a grande quantidade de étimos que temos em suas obras indianistas.

Ratificando nosso zelo em relação à Etimologia Ficcional Contextual de Alencar, trouxemos para procedermos às análises quatro dicionários de língua indígena, dentre eles o de GD, que foi publicado em período anterior à publicação dos romances de Alencar. Trouxemos também três dicionários de língua portuguesa, também anteriores às publicações de Alencar e, como complemento aos dicionários, o *Corpus do Português* (DAVIES, 2016) para consulta dos vocábulos. Esse *corpus* nos revelou, por meio da consulta, que quase a totalidade dos vocábulos indianistas consultados foram empregados por Alencar, pela primeira vez, em textos escritos no Português do Brasil.

Nosso terceiro objetivo:

i) Propor um desenho para elaboração de um Vocabulário com os vocábulos indianistas de Alencar, a ser disponibilizado *online*

No Capítulo 7 "Léxico indianista de José de Alencar: proposta para um Vocabulário *online*", nosso intento foi atingido, porque esboçamos uma proposta para esse Vocabulário. Preenchemos 17 fichas lexicográficas com dados que julgamos pertinentes para os consulentes dessa obra lexicográfica. Em seguida, apresentamos os verbetes baseados nas informações dispostas nas fichas.

Entendemos que a elaboração de uma obra lexicográfica para disponibilização *online*, requer um tratamento especializado, desenvolvido entre o lexicógrafo e o profissional da informática, para que o produto fique disponível com as características de consulta necessárias. Esse vocabulário é viável, na medida em que o leitor terá maior agilidade na busca pelo vocábulo desejado, se puder pesquisá-lo em uma fonte *online*. Isso também porque o perfil do leitor está se adaptando à utilização de dispositivos portáteis de leitura; sendo assim, é conveniente uma obra lexicográfica também disponível em formato eletrônico.

O Vocabulário com os vocábulos indianistas servirá para subsidiar o usuário nessas situações de leitura, além de possibilitar o entendimento de palavras que saíram dos romances para nomear pessoas como **Iracema**, **Moacir**, **Jurandir**, **Ubirajara**; espécies da flora e da fauna como **abacaxi, capim, cipó, gambá, tatu, juriti**, dentre outros. Assim também o léxico que compõe os demais campos semânticos faz parte da língua portuguesa do Brasil. Corroboramos com o que afirmou Gladstone Chaves de Melo (1971)

> Não deixa de ser muito significativo o alto número de vozes, que o tupi legou ao português do Brasil. Esse vocabulário novo reflete o nosso meio com seus pertences e suas riquezas, os componentes da nossa paisagem, as nossas coisas, a nossa vida, enfim. É a repercussão, na língua, da herança que o índio deixou no sangue e na cultura brasileira, é um sinal e uma presença do nosso passado, aventureiro e heroico, nas Estrada e Bandeiras(...). (MELO, 1971, p. 45)

Ressaltamos que mantivemos os antropônimos e os topônimos entre os vocábulos em análise, tanto na Etimologia Ficcional Contextual, quanto na composição das fichas lexicográficas. Essa decisão foi pautada no fato de que Alencar os utiliza com um propósito de imprimir características específicas, relacionadas ao caráter ou ao comportamento dos personagens, por meio da escolha do nome. Esse é um ato criativo do autor com base no ambiente nas características das pessoas e dos ambientes em que vivem. Ratificamos em Guérios (1973), entre os povos denominados primitivos, o nome da pessoa e do lugar eram inseparáveis de suas características.

Alencar conhecia e reconhecia a riqueza do Português utilizado no Brasil em relação ao Português de Portugal e como muito apregoado, trabalhou em defesa de uma língua genuinamente brasileira capaz de representar a nação em suas peculiaridades e, ao mesmo tempo, em suas diferenças de Portugal.

Estudar Alencar e o seu léxico indianista nos autoriza a dizer que o universo que envolve a vida dos indígenas também foi um elemento explorado pelo autor, que buscou no léxico a melhor forma de representar esse povo de cuja nação o autor ambicionava liberdade linguística.

A descrição da vida dos indígenas em todas as suas nuances adensa os seus romances tornando-os símbolos de uma tarefa empreendida com o propósito de valorizar o povo brasileiro e, para isso, partiu, inicialmente, da língua dos indígenas que, em sua essência, projeta o povo verdadeiramente brasileiro.

Assim, esta pesquisa buscou contribuir para o estudo do léxico indianista de Alencar, tanto no que passamos a denominar Etimologia Ficcional Contextual, como também pela proposta de elaboração do Vocabulário eletrônico com os termos indianistas, a ser disponibilizado *online*. Apesar das análises etimológicas dos étimos de Alencar apresentadas nesta tese, entendemos que não esgotamos as possibilidades de estudo sobre esse tema.

Deixamos aberto o campo de pesquisa para que, quem sabe, eu mesma ou outros pesquisadores possam dar continuidade aos estudos da riqueza lexical de Alencar.

Com base no exposto, acreditamos ter cumprido os objetivos propostos; contudo reconhecemos as limitações de nossa investigação, as quais se apresentam como novos caminhos abertos para investigações futuras.

Além das contribuições apresentadas anteriormente e, com relação à elaboração do Vocabulário com o léxico indianista *online*, para a disponibilização dessa obra lexicográfica com os dados apurados, será necessário um planejamento técnico, a fim de adequar as informações em relação aos dados a serem disponibilizados. Ademais, será necessário também um aprofundamento do projeto do Vocabulário indianista de Alencar para disponibilização *online*. Os procedimentos estabelecidos e as análises desenvolvidas neste estudo serão a base para a concretização desse projeto.

Para além desses propósitos, o Vocabulário *online* poderá contribuir para satisfazer o desejo de cultura do leitor que se vê, às vezes, assoberbado com tão grande número de vocábulos indígenas inseridos nas obras de Alencar.

Ao desenvolver uma pesquisa para escrevermos uma tese, podemos esbarrar em algumas dificuldades durante o percurso. No meu caso, diante de uma situação que parecia indicar o meu abandono do curso, quando restava um ano e seis meses para o término do Doutorado, considerando o prazo para defesa, uma nova possibilidade surgiu. Adotando outra linha de pesquisa, radicalmente oposta ao que vinha desenvolvendo, mudei meus passos para outra direção. Porém, com a mudança, a satisfação encontrada em pesquisar o léxico de Alencar aliada à tranquilidade em relação às orientações recebidas, tornaram possível desenvolver esta pesquisa e finalizar esta tese em um ano e dois meses.

Tomo a liberdade de finalizar esta tese, contrariando Alencar, ao dizer que “tudo passa sobre a Terra”:

- Não, Alencar! Nem tudo passa sobre a terra!

REFERÊNCIAS

ABBADE, C. M. S. Ó pai Ó e outras particularidades do léxico baiano. **Cadernos do CNLF**, vol. XII, no. 09. Disponível em: < http://www.filologia.org.br/xiicnlf/09/11.pdf Acesso em: 15 abr. 2017.

_____. A Lexicologia e a teoria dos campos lexicais. **Anais do VX Congresso Nacional de Linguística e Filologia.** v. XV, n. 5. Rio de Janeiro: CiFFil, 2011. p. 1332-1343. Disponível em: www.filologia.org.br/xv_cnlf/tomo_2/105.pdf. Acesso em: 15 abr. 2017.

ABREU, M. M. **Ao pé da página**: a dupla narrativa em José de Alencar. Campinas, SP: Mercado das Letras, 2011.

AIRES DE CASAL, M. 1754?-1821?. **Corografia brasílica ou Relação histórico-geográfica do Reino do Brasil**. Belo Horizonte: Ed. Itatiaia, 1976.

_____. **Corografia Brasílica.** Tomo 1. Rio de Janeiro: Imprensa Nacional, 1945.

ALENCAR, J. **Iracema**. Ed. do centenário. Rio de Janeiro: Livraria José Olympio, 1965.

_____. **O Guarani**. Tomo 1º. Rio de Janeiro: Livraria José Olympio, 1951.

_____. **O Guarani**. Tomo 2º. Rio de Janeiro: Libraria José Olympio, 1951.

_____. **Ubirajara**. Rio de Janeiro: Edições de Ouro. s/d

AMARAL, E. T. R. **Nomes próprios**: análise de antropônimos do espanhol escrito. 2008. Tese (Doutorado em Língua Espanhola e Literaturas Espanhola e Hispano-Americana). São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas da USPR, 2008.

ANTUNES, I. **Território das palavras**: estudo do léxico em sala de aula. São Paulo: Parábola Editorial, 2012

ÁVILA, M. V. D. **O léxico indianista em José de Alencar**: uma análise parcelar. 2004. 253 p. Dissertação de Mestrado (Mestrado em Estudos Linguísticos) – Uberlândia: Programa de Pós-graduação em Estudos Linguísticos – UFU, 2004.

ÁVILA, M.V.D; MATINS, E. S. O léxico indianista em José de Alencar. **Revista Vertentes**. São João Del Rei: UFSJ, 2008, p. 233-245, n. 32.

AZEREDO, José Carlos de. **Fundamentos de Gramática do Português**. 3ª ed. Rio de Janeiro: Jorge Zahar, 2004.

BARBOSA, M. A. Lexicologia, lexicografia, terminologia, terminografia: objeto, métodos, campos de atuação e de cooperação. In**: Anais dos Seminários do Grupo de Estudos Linguísticos de São Paulo** – *GEL*, 39, Franca: UNIFRAN, 1991, p. 182— 189.

____. Da neologia à neoloiga na literatura. In: OLIVEIRA, A. M. P. P. ; NEGRI, A. (org.) **As Ciências do Léxico**: Lexicologia, Lexicografia, Terminologia. 2 ed. Campo Granda: Ed. UFMS, 2001.

_____. Estrutura e formação do conceito nas línguas especializadas: tratamento terminológico e lexicográfico. **Rev. Bras. Linguist. Apl**., Belo Horizonte, v. 4, n. 1, p. 55-86, 2004, Disponível em: <www.scielo.br/pdf/rbla/v4n1/05.pdf>. Acesso em 05 abr. 2018. DOI: http://dx.doi.org./10.1590/S1984-63982004000100006.

BARROS, L. A. Aspectos epistemológicos e perspectivas científicias da terminologia. **Cien. Cult.** São Paulo, v. 58, n. 2, jun. 2006. Disponível em: http://cienciacultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200011&lng=en&nrm=iso. Acesso em: 14 out. 2017.

BATAR, P.; DeCESARIS, J. **De Lexicografia**. Barcelona, I'Iula: 2004.

BEIVIDAS, W. A teoria da linguagem de Hjelmslev: uma epistemologia imanente do conhecimento. **Estudos Semióticos**. São Paulo: USP, vol. 11, no. 1, p. 1-10.Olá Disponível em: <www.revista.usp.br/esse/article/view/103769/103467>. Acesso em: 07 jun. 2018.

BERBER SARDINHA, T. **Linguística de *Corpus***. Barueri, SP: Manole, 2004.

_____Pesquisa em Linguística de *Corpus* com *WordSmith Tools*. Campinas, SP: Mercado das Letras, 2009.

_____Linguística de *corpus*: histórico e problemática. Revista D.E.L.T.A., v. 16, n. 2, 2000, p. 323-367.

BIDERMAN, M. T. C. **Dimensões da palavra**. Filologia e linguística portuguesa. N. 2, 1998, p. 81-118.

_____Teoria Linguística: teoria lexical e computacional. São Paulo: Martins Fontes, 2001.

_____Léxico e vocabulário fundamental. Alfa, São Paulo, 40: 27-46, 1996.

_____A face quantitativa da linguagem: um dicionário de frequência do Português. Alfa, São Paulo, 42: 161-181, 1998.

_____. **Usando o WordSmith Tools na investigação da linguagem**. *DIRECT Papers*, n. 40. São Paulo: LAEL/PUC, 1999.

BLUTEAU, R. **Vocabulario Portuguez & Latino.** Coimbra: Collegio das Artes da Companhia de Jesus, 1712-1728.

BORBA, F. S. **Organização de Dicionários**: uma introdução à lexicografia. São Paulo: Ed. UNESP, 2003.

BOSI, A. **História concisa da literatura brasileira**.36ª. ed. São Paulo: Cultrix, 1999.

CÂMARA JÚNIOR, J. M. **Dispersos**. Rio de Janeiro: Fundação Getúlio Vargas, 1972.

_____. J. M. **Introdução às línguas indígenas brasileiras**. 2 ed. Rio de Janeiro: Livraria Acadêmica, 1965.

CANÇADO, M. **Manual de Semântica**: noções básicas e exercícios. Belo Horizonte: Editora UFMG, 2005.

_____. **Manual de semântica**: noções básicas e exercícios. Belo Horizonte: Editora UFMG, 2008.

CARDOSO, S. A. F. **Termosteo:** a elaboração de vocabulários monolíngues de termos da Teologia em um estudo conduzido por corpus. 2017. 315 f. Tese (Doutorado em Estudos Linguísticos) -Programa de Pós-Graduação em Estudos Linguísticos, Universidade Federal de Uberlândia, Uberlândia, 2017.

CARNEIRO, R. M. O. **Discurso literário de fantasia infanto-juvenil: proposta de descrição terminológica direcionada por *corpus***. 2016. 281 f. Dissertação (Mestrado em Estudos Linguísticos) - Programa de Pós-Graduação em Estudos Linguísticos, Universidade Federal de Uberlândia, Uberlândia, 2016.

CARVALHINHOS, P. J. As origens dos nomes de pessoas. **Domínios da Lingu@gem**. Ano1, no 1, 2007. Disponível em: <www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/11401/6686>. Acesso em: 08 jun. 2018.

CASARES, J. **Introducción a la Lexicografía moderna**. 3 ed. Madrid: Raycar, 1992.

CASTELO, J. A. Iracema e o indianismo de Alencar. In: PROENÇA, M. C. (org.). **Iracema**: José de Alencar. São Paulo: Edusp, 1979. p. 270-280.

COSTA, M. I. P. **Terminologia jurídico-policial**: proposta de elaboração de um glossário eletrônico. 2014. 287f. Tese (Doutorado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2014.

COSTA, E. O.; SALES, G. M. A. O léxico do vestuário feminino no século XIX: o frolido de sedas na narrativa de José de Alencar. **Revista do Programa de Pós-Graduação em Letras da Universidade do Passo Fundo**. v. 11, n. 2 – p. 357-375. Jul./dez. 2015.

COUTO, H. H. **Onomasiologia e semasiologia revisitadas pela ecolinguística**. Revista de Estudos da Linguagem. Vol. 20, no 2. (2012). Disponível em: http://periodicos.letras.ufmg.br/index.php/relin/article/view/2748/2703>. Acesso em: 14 out. 2017.

CUNHA, A. G. **Dicionário histórico das palavras de origem tupi**. 4ª ed. São Paulo: Companhia Melhoramentos; Brasília: Universidade de Brasília, 1998.

DAVIES, M. ***Corpus* do Português**. 2016. Disponível em: <www.corpusdoportugues.org.> Acesso em: 05 ago. 2017.

DICK, M. V. P. A. Métodos e questões terminológicas na onomástica. Estudo de caso: o atlas toponímico do estado de São Paulo. **Revista Investigações, Linguística e Teoria Literária**, Recife, v. 9, p. 119-149, março, 1999.

DRAGO, I.; VIEIRA, A. B.; SILVA, A. P. C. **Caracterização dos arquivos armazenados no Dropbox**, 2013. Disponível em: < http://www.4nerd.net/2013/2013_sbrcWP2P_drop.pdf> Acessoem: 16 set. 2017.

DUBOIS, J. et al. **Dicionário de Linguística**. 14 ed. São Paulo: Cultrix, 2007.

_____& DUBOIS, C. **Introduction à la lexicographie:** le dictionnaire. Paris: librairie Larousse, 1971.

ECKERT, K.; RÖHRIG, M. antAntroponímia ficcional: o caso de Ubirajara, de José de Alencar. **Revista GTLex**. Uberlândia, v 2 n. 1, jul./dez. 2016. Disponível em: <http://www.seer.ufu.br/index.php/GTLex/article/view/37831/20533>. Acesso em: 30 maio 2018. DOI: 10.14393/Lex3-v2n1a2016-7.

FAULSTICH, E. L. J. **Lexicologia, a linguagem do noticiário policial**. Brasília: Horizonte, 1980.

FIORIN, J. L. Considerações em torno do projeto de Lei n. 1676/99. In: FARACO, C. A. (Org.) **Estrangeirismos*:*** guerras em torno da língua. São Paulo: Parábola, 2001.

FROMM. G. Obras lexicográficas e terminológicas: definições. **Revista Factus**. v. 1, n. 2, p. 139-147, 2004.

_____. **VoTec: a construção de vocabulários eletrônicos para aprendizes de tradução**. 2007. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês) – Departamento de Letras Modernas, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

GARCÍA PALACIOS, J. El artículo lexicográfico en el dicionário de especialidade. In: AHUMADA, I. (Ed.) **Diccionarios y lenguas de especialidade**. V Seminario de Lexicografia Hispánica. Jaén, 21-23 de novembro de 2001. Jaén: Universidad de Jaén, Servicio de Publicaciones de la Universidad de Jaén, 2002, p. 21-47.

GODOI, E. **O vocabulário indianista e ideológico de José de Alencar**. Linguagem – Estudos de Pesquisas, Catalão, v. 8-9, p. 84-100, 2006.

GONÇALVES DIAS. A. **Dicionário da língua Tupy**. Lipsia: F. A. Brockhaus, 1858.

GUIRRAUD, P. **A semântica**. Tradução e adaptação de Maria Elisa Mascarenhas. São Paulo: Difusão europeia do livro, 1972.

HAENSCH, G. Tipología de las obras lexicográficas. In: HAENSCH, G. et al. **La Lexicografia:** De la linguística teórica e la lexicografia práctica. Madrid: Credos, 1982, p. 95-187.

HOUAISS, A. **Dicionário eletrônico Houaiss da língua portuguesa**. Versão 3.0, 2009.

ISQUERDO. A. N. A Toponímia como signo de representação de uma realidade. Fronteira – **Revista de História** (UFMS), Campo Grande-MS, v. 1, n. 2, p. 27-46, jul./dez. 1997.

KATZ, J. J.; FODOR, J. A. Estrutura de uma teoria semântica. In: LOBATO, L. M. P. **A semântica na linguística moderna**: o léxico. Rio de Janeiro: F. Alves, 1977, 388 p.

KRIEGER, M. G.; FINATTO, M. J. B. **Introdução à terminologia**: teoria e prática. São Paulo: Contexto, 2004.

LEFFA, V. J. **O uso de dicionários on-line na compreensão de textos em língua estrangeira**. Trabalho apresentado no VI Congresso Brasileiro de Linguística Aplica. Belo Horizonte: UFMG, 7- 11 de outubro de 2001. p. 39 (resumo). Disponível em: <http://www.leffa.pro.br/textos/trabalhos/dicionario.pdf>. Acesso em: 25 maio 2018.

_____**O dicionário eletrônico na construção do sentido em língua estrangeira**. Cadernos de tradução, Florianópolis, n. 18, p. 319-340, 2006. Disponível em: < http://www.leffa.pro.br/textos/trabalhos/dic_eletronic.pdf>. Acesso em: 25 maio 2018.

MACIEL, A. M. B. Terminologia, linguagem de especialidade e dicionários. In: KRIEGER, M. G. e MACIEL, A. M. B. (Org.) **Temas e Terminologia**. Porto Alegre/São Paulo: Ed. Universidade/UFRGS/Humanistas/USP, 2001.

MARCO, V. **A perda das ilusões**: o romance histórico de José de Alencar. Campinas: Editora da Unicamp, 1993, p.99.

MARTINS, S. C. **Proposta de uma base de conhecimento multilíngue *on-line* de expressões cromáticas da *Fauna* e da *Flora***. 2017. 432f. Tese (Doutorado em Estudos Linguísticos) Pós-Graduação em Estudos Linguísticos, Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

MELO, G. C. **A Língua do Brasil**. Rio de Janeiro: Fundação Getúlio Vargas, 1971.

MEYER, A. Alencar. In: PROENÇA, M. C. (org.). **Iracema**: José de Alencar. São Paulo: Edusp, 1979. p. 185-204.

MIRANDA, F. B. LA ETIMOLOGÍA EN EL DICCIONARIO DE LA LENGUA**. Revista Letras**, [S.l.], v. 64, dez. 2004. ISSN 2236-0999. Disponível em: <https://revistas.ufpr.br/letras/article/view/2976/2404>. Acesso em: 28 maio 2018. doi:http://dx.doi.org/10.5380/rel.v64i0.2976.

MORAES PINTO, M. C. **A vida Selvagem**: paralelo entre Chateaubriand e Alencar. São Paulo: Annablume, 1995.

NAVARRO, E. A. **Dicionário de tupi antigo**: a língua indígena clássica do Brasil. São Paulo: Global, 2013.

NOVODVORSKI, A.; BOCORNY FINATTO, M. J. Linguística de *Corpus* no Brasil: uma aventura mais do que adequada. **Letras & Letras**, [S.l.], v. 30, n. 2, p. 7-16, dez. 2014. ISSN 1981-5239. Disponível em: <http://www.seer.ufu.br/index.php/letraseletras/article/view/28516>. Acesso em: 09 abr. 2017. doi: http://dx.doi.org/10.14393/LL60-v30n2a2014-1.

NOVODVORSKI, A.; HERRERA, A. M. F. Conversaciones con un lingüista de *corpus*: Professor Dr. Giovanni Parodi. **Letras & Letras**, Vol. 30, Nº 2. Edufu: Uberlândia, 2014, p. 452- 466. PARODI, G. Lingüística de *Corpus*: de la teoría a la empiria. Madrid: Iberoamericana / Vervuert,

NOVODVORSKI, A. **Estilo das traduções de Sergio Molina de obras de Ernesto Sabato**: um estudo de corpora paralelo Espanhol/Português. 2013. 259f. Tese (Doutorado) - Curso de Doutorado em Estudos Linguísticos, Universidade Federal de Minas Gerais, Belo Horizonte.

PAGNAN, C. L. Alencar e Machado: Leituras e indianismo. **Revista de Estudos Literários.** Disponível: <www.uel.br/pos/letras/terraroxa>. Acesso em: 28 maio 2018.

PARODI, G. **Linguística de *Corpus***: de la teoría a la empiria. Madrid; Frankfurt: Iberoamericana – Vervuert, 2008.

PAVEL, S.; NOLET, D. **Manual de terminologia**. Tradução de Enilde Faulstich. Canadá: Departamento de Tradução, 2002.

PEREIRA, M. T. G. De Saussure, de outras contribuições, de ocorrências linguísticas: a relevância da etimologia popular. **Revista Matraga.** Vol 21, n. 34 (2014). Disponível em: < http://www.e-publicacoes.uerj.br/index.php/matraga/article/view/17511>. Acesso em: 28 maio 2018.

PINTO, L. M. S. **Dicionário da Lingua Brasileira.** Ouro Preto: Typographia de Siva, 1832.

POTTIER, B. **Estruturas Linguísticas do Português**. Tradução de Alberto Audubert e Cidimar Teodoro Pais. São Paulo, Difusão Europeia do livro, 1972.

PRETI, D. **Sociolinguística, os níveis da fala**. 3 ed. São Paulo: Companhia Editorial Nacional, 1977.

PROENÇA, M.C. **Estudos Literários**. Rio de Janeiro: Livraria José Olympio, 1969.

QUEIROZ, S. R. S. **O vocabulário alencariano de O Sertanejo:** uma análise léxico-semântica. 2006. 357 p. Dissertação de Mestrado (Mestrado em Estudos Linguísticos) – Programa de Pós-graduação em Estudos Linguísticos – PPGEL, Universidade Federal de Uberlândia, 2006.

QUEIROZ, R. José de Alencar. In: ALENCAR, J. **Iracema**. Ed. do centenário. Rio de Janeiro: Livraria José Olympio, 1965.

RAMOS, I. P. **Ubirajara**: ficção e fricções alencarianas. Belo Horizonte, v. 11, p. 59-63, dez. 2007. Disponível em: <http://www.periodicos.letras.ufmg.br. Acesso em: 02 jul. 2017.

REY-DEBOVE, J. **Léxico e dicionário**. Tradução de Clóvis Barleta de Morais. Alfa, São Paulo, 28(supl.):45-69, 1984.

ROCHA, B. Introdução Biográfica. In: ALENCAR, J. **Iracema**. Ed. do centenário. Rio de Janeiro: Livraria José Olympio, 1965.

SAMPAIO. T. **O tupi na geografia nacional**. São Paulo: Casa Eclectica, 1901.

SANTIAGO, S. Roteiro para uma leitura intertextual de Ubirajara. In: ALENCAR, José de. **Ubirajara**. 8ª ed. São Paulo: Ática, 1984. (Texto introdutório do romance Ubirajara)

SAUSSURE, F. **Curso de Linguística Geral**. Cultrix: São Paulo, 1975.

SCHIERHOLZ, S. J. Lexicografia de Especialidade e Terminografia. Tradução de Leonardo Zilio. In: **Cadernos de Tradução**. Corpus, Corpora e Dicionários. Porto Alegre, no 30, jan-jun, 2012, p. 1-64.

SCOTT, M. Problems in investigating keyness, or clearing the undergrowth and marking out trails... In: BONDI, M. SCOTT M. (Eds.) **Keyness int Texts.** Amsterdam/Philadelphia: John Benjamins Publishing Company, 2010. P. 43-57. Disponível em: https://benjamins.com/catalog/scl.41.04sco. Acesso em 25 maio 2018. Doi: https://doi.org/10.1075/scl.41.04sco.

SILVA, A, M. **Diccionario da Lingua Portugueza**. Lisboa: Officina de Simão Thaddeo Ferreira, 1789.

SILVA, E. B. **Proposta de um dicionário eletrônico terminológico onomasiológico bilíngue inglês-português no domínio das redes neurais artificiais**. 2008, 143 p. Dissertação de Mestrado (Mestrado em Análise Linguística) Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista, 2008.

SILVIANO, S. **Ora (direis) puxar conversa**!: ensaios literários. Belo Horizonte: Editora UFMG, 2006. 381 p.

SCHREINER, C. **Edição de Documentos e Estudo do Vocabulário do Charque na Região Sul do Brasil:** contribuição à história do português brasileiro. 2012. 504 p. Tese de Doutorado. (Doutorado em Letras) Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2012.

SCHWAMBORN, I. **O Guarani era um Tupi?** – Sobre os Romances Indianistas O Guarani, Iracema, Ubirajara de José de Alencar. Tradução: Carlos Almeida Pereira. Fortaleza: Casa de José de Alencar/Programa Editorial, 1998.

SCOTT, M. **WordSmith Tools**. Versão 6. Liverpool: Lexical Analysis Software, 2012.

STUBBS, M. **Text and corpus analysis**. Cambridge, Massachusetts: Blavjwell Publishers, 1996.

TÁVORA, F. **Cartas a Cincinato**. Estudos críticos de Semprônio. Pernambuco: J.W. Medeiros.

TIBIRIÇÁ. L. C. **Dicionário tupi-português**. Liberdde/SP: Traço Editora, 1984.

TOGNINI-BONELLI, E. **Corpus Linguistic at Work**. Amsterdam: John Benjamins, 2001.

TRAVAGLIA, L. C. O ensino de vocabulário e sua importância. In: BASTOS, N. B.(org.) **História, Cultura e Sociedade**. São Paulo: EDUC: IP-PUCSP, 2016.

VARNHAGEM. F. A. **História da Independência do Brasil**. 6 ed. São Paulo, Melhoramentos, 1972.

História Geral do Brasil. 1816-1878. 10 ed. Belo Horizonte: Ed. Itatiaia; São Paulo: Ed. Universidade de São Paulo, 1981.

VASCONCELOS, A. F. "**A verdade dispensa a verossimilhança**": o fato e a ficção no romance histórico As Minas de Prata de José de Alencar. 2011. 164 p. Dissertação de Mestrado (Mestrado em Letras) Universidade Federal do Ceará, 2011.

VIARO, M. E. **Etimologia**. São Paulo: Contexto, 2014.

VILELA, M. **Ensino da Língua Portuguesa**: Léxico, Dicionário, Gramática. Coimbra: Almedina, 1995.

Estruturas léxicas do Português. Coimbra: Livraria Almedina, 1979.

VILLALVA, A., SILVESTRE, J. P. **Introdução ao estudo do léxico:** descrição e análise do Português. Petrópolis, RJ: Vozes, 2014.

WELKER, H. A. **Dicionários** – uma pequena introdução à Lexicografia. 2 ed. Brasília: Thesaurus, 2004.

APÊNDICES

APÊNDICE A – Quadro comparativo entre as 20 primeiras palavras extraídas por meio da ferramenta KeyWords, do WST. *Corpus* de Estudo com os *corpora* de referência.

Corpus de estudo	*Corpus* de Referência Alencar	*Corpus* de Referência AcadTeses	*Corpus* de Referência Lácio-Ref	*Corpus* de Referência Nov
Peri	X	X	X	X
Cecília	X	X	X	X
Guerreiro	X	X	X	X
Índio	X	X	X	X
Guerreiros	X	X	X	X
Álvaro	X	X	X	X
Iracema	X	X	X	X
Loredano	X	X	X	X
Chefe	X		X	X
Ubirajara	X	X	X	X
Antônio	X		X	X
Italiano	X			X
Aventureiros	X	X	X	X
Cabana	X	X	X	X
Mariz	X	X	X	X
Itaquê	X	X		
Virgem	X	X	X	X
Poti	X	X		
Taba	X	X		
Araci	X			
olhos		X	X	X
fidalgo		X	X	X
menina		X	X	
Isabel		X	X	X
senhora			X	X

APÊNDICE B – Quadro comparativo entre as 20 primeiras palavras com frequência zero extraídas por meio da ferramenta KeyWords, do WST. *Corpus* de Estudo com os *corpora* de referência.

Corpus de estudo	*Corpus* de Referência Alencar	*Corpus* de Referência AcadTeses	*Corpus* de Referência Lácio-Ref	*Corpus* de Referência Nov
Loredano	X		X	X
Ubirajara	X			
Itaquê	X	X		X
Poti	X		X	X
Pojucã	X	X		X
Jurandir	X			X
Araquém	X	X	X	X
Tocantim	X	X		X
Jaguarê	X	X		X
Araguaias	X	X		X
Jandira	X			X
Araguaia	X			
Tabajaras	X		X	
Caubi	X	X		
Irapuã	X			
Soeiro	X			
Simões	X			
Pitiguaras	X	X	X	
Distancia	X			
Pitiguara	X	X	X	
Ergueu-se		X		
Camacã		X		
Lembrou-se		X		
Dirigiu-se		X		
Moacaras		X		
Ouviu-se		X		
Canicrã		X		
Abarés		X		
Tinha-se		X		
Tornou-se		X		
aimoré			X	X
Lauriana			X	
Sabeis			X	
quereis			X	
murmurou			X	
jacaúna			X	
relva			X	
sois			X	
clavina			X	
Abarê			X	
Peri				X

Guerreiros				X
cabana				X
Martim				X
Tupã				X
pajé				X
esposo				X
seta				X

APÊNDICE C – Lista das 367 palavras com indicação das obras em que são utilizadas

Palavra	obras	Palavra	obras	Palavra	obras	Palavra	obras	Palavra		obras
abacaxis	U	abaeté	I	abaré	U	abati	I	açaí		U
acajás	U	acaraú	I	acaris	G	acarú	I	acauã		I
aguapés	G	aimoré	G	aimorés	G	aimóres	G	aipim		U
airi	U	anajê	I	ananás	IUG	andira	I	andiroba		I
angico	I	anhangá	IU	anum	U	apodi	I	ará		I
araçá	I	araçás	G	aracati	I	araci	U	araçóia		IU
araguaia	U	araguaias	U	araquém	I	arara	IU	ararê		G
araribás	G	araribóia	U	araruna	U	aratanha	I	aratuba		U
aresqui	U	arirama	U	ariranha	I	aroeira	G	aroeiras		U
arroio	G	bambu	I	ati	I	atiati	GI	balaio		U
banana	U	bananas	U	Bananeiras	IU	batuireté	I	biaribi		U
biribá	G	boicininga	I	boitatá		boré	I	borés		U
cabuíba	G	cacique	G	caiçara	IU	caiçaras	U	caiporas		U
caitetus	U	caititus	I	cutia	IGU	cajazeira	U	caju		IGU
cajueiro	GI	cajueiros	I	cajus	I	camacã	U	camoropim		I
camorupim	I	campeiro	GU	campinas	IU	camuci	G	Camucim		IU
camucins	IU	canicrã	U	caninana	I	capim	G	capivara		IGU
capoeira	I	caraíba	U	caramuru	G	caramurus	U	caraúba		U
carcará	U	carimã	I	carioba	I	carnaúba	I	catuíba		G
cauã	IG	cauatá	U	caubi	I	cauim	IGU	ceará		I
graúna	I	cearense	I	ceci	G	choça	G	cipó		IGU
cipós	G	clavina	G	coati	I	coatiabo	I	copaíba		I
jetaí	I	corisco	U	coruja	G	crajuá	U	crajuru		IU
crauatá	U	craúba	U	crautá	IU	craviri	U	crebã		U
croás	U	cuandu	I	cumari	U	cumbucas	U	cupim		I
curumim	U	embaíba	GU	emboabas	I	gará	I	goaná		I
goiamum	I	goitacá	G	goitacás	G	guabiroba	I	guaiúba		I
guanumbi	GU	guará	IU	guaraciabas	I	guaraná	U	guarani		G
guaranis	G	guaribu	U	guaximas	G	guaxinim	U	iara		G
ibiapaba	I	ibiapina	I	icó	I	igaçaba	I	igaçabas		IU
igapê	U	igara	I	igaras	IG	iguape	I	imbé		U
imbu	I	inhuma	IU	intanha	I	Inúbia	IGU	inúbias		G
ipu	I	ipus	I	iracema	I	irapuã	I	irara		IG
irerê	U	itaoca	IU	itaquê	U	itaquêe	U	jaburu		I
jabuti	U	jacamim	U	jaçan	U	jaçanã	IU	jaçanãs		IG
jacarandá	IGU	jacarandás	G	jacaré	GU	jacarecanga	I	jacareí		I
jacaúna	I	jaci	I	jacobina	G	jacus	I	jaguar		IGU
jaguaraçu	1	jaguarê	U	jaguaribe	I	jandaias	I	jandira		U
japi	I	japim	I	jararaca	G	jataí	I	jati		I
jatobá	IU	jatobás	U	javari	U	jenipapo	I	jequiriti		U

jequis	U	jereraú	I	jetica	I	jibóia	IGU	jibóias		G
jirau	IU	juazeiros	I	juçara	IU	juquiri	U	jurandir		U
jurema	I	juriti	I	jurupari	I	juruti	GU	jutaí		U
jutorib	U	macana	U	mairi	I	majé	U	majoí		I
manacá	IU	monguba	IU	manava	I	mandioca	IU	mangaba		I
mangabeira	U	maniva	IU	marabá	U	maracá	IGU	maracajá		I
maracás	G	maracatim	I	maracujá	IU	maranduba	U	maranguab		I
mearim	I	membi	I	meruoca	I	mingau	I	moacaras		U
moacir	I	mocejana	I	mocoribe	I	moquém	IU	morubixaba		U
muçurana	GU	mundaú	I	murinhém	U	muriti	I	muritiapuá		I
muritis	I	mururê	U	mutucas	U	nambu	IU	nandu		U
nhengaçara	U	nhengaçaras	U	nhengaíba	U	oca	IU	ocara		IU
ogib	U	oitibó	IG	oticica	I	ouricuri	G	paca		I
pacas	IG	pacatuba	I	pacoti	I	pahã	U	pajé		IU
pajés	GU	papagaios	I	paquequer	G	pará	U	paraíba		IG
quati	I	parnaíba	I	pequiá	G	peri	G	piaçaba		U
piau	I	piracém	I	pirajá	U	piranhas	I	pirapora		I
pirijá	U	pirogas	I	pitanga	I	pitiguara	I	pitiguaras		I
pocema	IU	pojucã	U	porongaba	I	poraquê	I	pororoca		U
potengi	I	poti	I	potiguara	I	potiguaras	I	quixeramobim		I
sabiá	IGU	sabiás	G	sagui	G	saí	I	saixê		G
sapé	IG	sapiranga	I	sapopema	IU	sapoti	U	sapucaia		GU
saúva	U	saúvas	U	sopé	G	sucuri	IU	taba		IGU
tabajara	I	tabajaras	I	tabas	IU	taboca	G	tacape		IGU
tacapes	IGU	taioba	I	tamandaré	G	tamanduá	I	tamoios		U
tanatinga	I	tangapema	G	tapir	IGU	tapuia	IU	tapuias		IU
tapuitinga	I	taquara	I	taquari	U	tatu	U	tauape		I
teiú	U	ticum	G	tocantim	U	tocantins	U	traíra		I
traíras	I	trairi	I	trocano	U	tubim	U	tucano		IGU
tucanos	U	tuins	I	tupã	IU	tupi	IU	tupinambá		I
tupinambás	IU	ubaia	IU	ubirajara	U	ubirajaras	U	ubiratã		IU
uiraçaba	IU	uiraçu	U	uraçá	I	urataí	G	uru		IU
uruburetama	I	urubus	IG	urus	U	Urutau	IG	uvaias		G
xingu	U	zabelê	IU							

Legenda: I – Iracema;
G – O Guarani;
U – Ubirajara;
IGU – três obras;
IG – Iracema e O Guarani;
IU – Iracema e Ubirajara;
GU – O Guarani e Ubirajara