Pontifícia Universidade Católica
do Rio de Janeiro

Mariana Lemos Muller

Espírito de Corpus – criação de um léxico bilíngue do Corpo de Fuzileiros Navais

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Letras/Estudos da Linguagem pelo Programa de Pós-Graduação em Estudos da Linguagem do Departamento de Letras da PUC-Rio.

Orientadora: Maria Cláudia de Freitas

Rio de Janeiro
Abril 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Mariana Lemos Muller

Espírito de Corpus – criação de um léxico bilíngue do Corpo de Fuzileiros Navais

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Estudos da Linguagem da PUC-Rio. Aprovada pela Comissão Examinadora abaixo:

Maria Cláudia de Freitas
Orientadora
Departamento de Letras – PUC-Rio

Janine Maria Mendonca Pimentel
UFRJ

Maria Jose Bocorny Finatto
UFRGS

Rio de Janeiro, 27 de abril de 2022.

Mariana Lemos Müller

Licenciada em Letras Português/Inglês pela Universidade Veiga de Almeida e pós-graduada em Técnicas, Práticas e Estudos da Tradução (Inglês – Português) pela PUC-Rio. Atua como tradutora de inglês desde 2006. Desde 2014, é Oficial RM2 do Quadro Técnico de Inglês da Marinha do Brasil e atua como instrutora e tradutora no Corpo de Fuzileiros Navais.

Ficha Catalográfica

Muller, Mariana Lemos

Espírito de Corpus : criação de um léxico bilíngue do Corpo de Fuzileiros Navais / Mariana Lemos Muller ; orientadora: Maria Cláudia de Freitas. – 2022.

124 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Letras, 2022.

Inclui bibliografia

1. Letras – Teses. 2. Corpo de Fuzileiros Navais. 3. Léxico. 4. Estudos da tradução baseados em Corpus. 5. Terminologia. 6. Linguística de Corpus. I. Freitas, Maria Cláudia de. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD: 400

Agradecimentos

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

Agradeço a minha família, que me apoiou em todos os momentos do curso e compreendeu a minha ausência. Em especial, ao meu esposo, o Capitão de Fragata (FN) Alexandre de Menezes Villarmosa, pela imensa contribuição como especialista do Corpo de Fuzileiros Navais, pelos longos debates sobre computação, terminologia, métricas etc. e pela imensa contribuição como programador na publicação do léxico na internet.

À Profa. Dra. Janine Pimentel, minha professora da Pós-Graduação em Técnicas e Estudos da Tradução, concluída em 2019, que me apresentou ao mundo da Terminologia e da Extração Automática de Termos com suas aulas fascinantes e, ainda, foi orientadora da minha monografia.

Aos meus professores do Mestrado, especialmente à minha orientadora, a Profa. Dra. Cláudia Freitas, que fez o impossível se tornar real; que me instigou a ir mais fundo nas investigações e a nunca desistir, mesmo quando parecia impossível, e que, por isso, a cheguei a lugares que nunca imaginei no âmbito da Linguística Computacional.

Ao Vice-Almirante (FN) Carlos Chagas, pelo imenso conhecimento em terminologia bilíngue do CFN e pela contribuição com a apresentação que gerou o corpus paralelo utilizado neste estudo.

Ao Contra-Almirante (FN) Cláudio Leite, que permitiu e compreendeu minha eventual ausência do trabalho a fim de escrever esta dissertação.

Ao Capitão de Mar e Guerra (FN) Luciano Dutra, pela contribuição em terminologia militar e pela valiosa indicação de material de referência.

Ao Capitão de Mar e Guerra (FN) Gonçalves Maia, o qual, ao permitir que eu colaborasse como tradutora e intérprete no Centro de Operações de Paz de Caráter Naval da Marinha do Brasil, possibilitou-me aplicar a Linguística de Corpus na criação dos primeiros glossários do CFN, aprimorando assim este estudo. Também contribuiu

sobremaneira nas minhas pesquisas sobre terminologia militar e naval ao permitir que eu cursasse o VI Estágio de Operações de Paz para Mulheres.

Ao Capitão de Fragata (FN) Maia, oficial de intercâmbio do CFN no USMC, que me auxiliou na revisão dos postos e graduações do léxico *Espírito de Corpus*.

Ao Capitão-Tenente (FN) Lourenço, por me fornecer rico material especializado imprescindível para a criação dos corpora utilizados neste estudo e pelo incentivo de sempre.

Resumo

Lemos Müller, Mariana; Freitas, Cláudia (Orientador). *Espírito de Corpus* – criação de um léxico bilíngue do Corpo de Fuzileiros Navais. Rio de Janeiro, 2022. 124 p. Dissertação de Mestrado**.** Programa de Pós-graduação em Estudos da Linguagem**,** Pontifícia Universidade Católica do Rio de Janeiro.

Este estudo apresenta uma pesquisa temática envolvendo Terminologia, Estudos de Tradução Baseados em Corpus, Terminologia Computacional e Semântica Lexical, e tem como objeto de estudo a área do Corpo de Fuzileiros Navais. O objetivo desta pesquisa foi de criar um material terminológico por meio de uma metodologia híbrida de extração de termos desenvolvida a partir de testes com ferramentas de Extração Automática de Termos (EAT). Assim, buscou-se solucionar tanto problemas tradutórios relacionados à subárea de estudo quanto à detecção e validação de candidatos a termos em um corpus. Primeiramente, foi realizado um estudo piloto com o objetivo de avaliar as ferramentas TermoStat Web 3.0 e AntConc 3.5.7. Após os testes por meio da análise de um corpus paralelo bilíngue, foram selecionadas as melhores condições identificadas para se obter uma metodologia eficaz de extração automática de termos aliada à análise humana. Em seguida, essa metodologia foi utilizada para a análise de um corpus bilíngue comparável. Os candidatos a termos extraídos foram então validados pelos critérios de Semântica Lexical propostos por L'Homme (2020) e, em seguida, foram detectados seus equivalentes terminológicos. Este estudo resultou na criação do léxico bilíngue *Espírito de Corpus*.

Palavras-chave

Corpo de Fuzileiros Navais; Léxico; Terminologia; Estudos da Tradução Baseados em Corpus; Terminologia Computacional; Semântica Lexical; Extração Automática de Termos; AntConc; TermoStat; USMC; CFN; Língua Inglesa; Língua Portuguesa.

Abstract

Lemos Müller, Mariana. ***Espírito de Corpus* – creation of a Marine Corps bilingual lexicon.** Rio de Janeiro, 2022. 124 p. Conclusion Dissertation of the Postgraduate Course in Language Studies, Pontifical Catholic University of Rio de Janeiro.

This study presents a thematic research in the Marine Corps area involving Terminology, Corpus-Based Translation Studies, Computational Terminology and Lexical Semantics. The objective of this research was to create a terminological material through a hybrid methodology of term extraction developed from tests with Automatic Term Extraction (ATE) tools. Thus, we sought to solve both translation problems related to the subarea of study and to the detection and validation of term candidates in a corpus. First, a pilot study was conducted aiming to analyze two tools – TermoStat Web 3.0 and AntConc 3.5.7. After the conduction of the tests through the analysis of a bilingual parallel corpus, the best conditions identified were selected to obtain an effective methodology of automatic extraction of terms allied to human analysis. Then, this methodology was used for the analysis of a comparable bilingual corpus. The term candidates automatically extracted were then validated by the Lexical Semantics criteria proposed by L'Homme (2020) and their translation equivalents were detected. This study resulted in the creation of the bilingual lexicon *Espírito de Corpus*.

Keywords

Marine Corps; Lexicon; Terminology; Corpus-based Translation Studies; Computational Linguistics; Lexical Semantics; Automatic Term Extraction; AntConc; TermoStat; USMC; CFN; English; Portuguese.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Quadros

Lista de Apêndices

1
Introdução

O Corpo de Fuzileiros Navais (CFN) da Marinha do Brasil é uma força militar anfíbia, ou seja, que atua no mar e em terra, cuja missão é prover a segurança do país no que se refere aos seus interesses navais, conforme BRASIL (2020a, p. 51):

> "Para assegurar sua capacidade de projeção de poder, a Marinha do Brasil possuirá, ainda, meios de Fuzileiros Navais, em permanente condição de pronto emprego para atuar em operações de guerra naval, em atividades de emprego de magnitude e permanência limitadas. A existência de tais meios é também essencial para a defesa dos arquipélagos e das ilhas oceânicas em águas jurisdicionais brasileiras, além de instalações navais e portuárias, e para a participação em operações internacionais de paz, em operações humanitárias e em apoio à política externa em qualquer região que configure cenário estratégico de interesse. Nas vias fluviais, serão fundamentais para assegurar o controle das margens durante as Operações Ribeirinhas. O Corpo de Fuzileiros Navais, força de caráter anfíbio e expedicionário por excelência, constitui-se em parcela do Conjugado Anfíbio da Marinha do Brasil."

A fim de entender melhor a subárea de conhecimento em que o CFN se insere (subordinada à área militar), é fundamental visualizar como a força se organiza e conhecer um pouco da sua história. Quando falamos em área militar, nesta pesquisa, as definições a que nos referimos são as de Oxford (2021): "1. relativo a guerra, a soldado e a Exército. 2. relativo às forças armadas (Marinha, Exército e Aeronáutica), à sua organização, às suas atividades".

No Brasil, nossas Forças Armadas se organizam da seguinte forma, conforme sumarizado na figura 1:

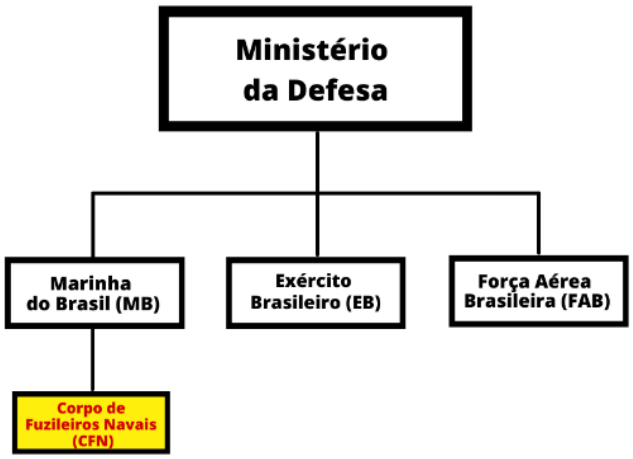


Figura 1 – Organograma das Forças Armadas Brasileiras. Fonte: elaborado pela autora com dados de Brasil, 2021A.

Observa-se na figura 1, portanto, que o CFN é uma força militar subordinada à MB que, por sua vez, está subordinada ao Ministério da Defesa, assim como as outras Forças Armadas existentes no país (EB e FAB).

A força anfíbia data de 1808, quando a Brigada Real da Marinha, vinda de Portugal, chegou ao Brasil acompanhando a Família Real lusitana, que tentava fugir de Napoleão Bonaparte. Segundo Medeiros e Oliveira (2013, p. 8), sua origem "remonta à criação da Brigada Real da Marinha de Portugal, por meio do alvará da D. Maria I, em 28 de agosto de 1797". O primeiro conflito em que atuou ocorreu em Caiena, na Guiana Francesa, a fim de impedir as tropas napoleônicas de adentrarem na região amazônica. Após tal ação, foi conquistado o atual território do Amapá (BRASIL, 2020b).

A denominação de Corpo de Fuzileiros Navais foi dada em 1932. Os Fuzileiros Navais atuaram ativamente em diversas batalhas ao longo da história do Brasil, em especial na Segunda Guerra Mundial, ao salvaguardar a costa brasileira em navios mercantes (BRASIL, Ib. Idib.).

Desde a sua criação, já se passaram 213 anos de história e tradição únicas, moldadas pelas características geográficas, socioeconômicas e culturais da nação brasileira, que demarcam as peculiaridades do CFN. Por isso, é difícil compará-lo com outras forças militares semelhantes pelo mundo, como o Corpo de Fuzileiros Navais dos Estados Unidos (*United States Marine Corps* - USMC), por exemplo, que é uma força independente da Marinha dos Estados Unidos da América (EUA) em diversos quesitos. Dessa forma, suas idiossincrasias causam normalmente problemas tradutórios ao se buscar equivalência em língua estrangeira entre os termos

que utiliza. A necessidade de um estudo da terminologia bilíngue (português e inglês) específica do CFN foi constatada pela autora ao longo de 7 anos servindo em um centro de instrução do CFN, o que motivou a criação do trabalho *Estudo terminológico bilíngue de termos militares da subárea do Corpo de Fuzileiros Navais* (MÜLLER, 2019).

Um exemplo de dificuldade de tradução comum proveniente das peculiaridades dessa força brasileira pode ser notado quanto à equivalência entre postos (patentes de oficiais) e graduações (patentes de praças) de forças estrangeiras semelhantes. Devido ao CFN ser subordinado à MB, as duas forças utilizam a mesma nomenclatura para suas patentes militares. A única diferença ocorre quanto ao termo *marinheiro*, utilizado para o militar mais moderno das praças apenas na MB, enquanto o CFN utiliza o termo *soldado* (BRASIL, 2021c). Tomemos como exemplo a relação de postos de oficiais da MB e, consequentemente, do CFN, mostrada na figura 2.

Figura 2 – Postos da MB. Fonte: Brasil, 2021b.

Na figura, podemos observar a nomenclatura utilizada para os postos da MB e do CFN, com grafia atualizada[1]. Porém, ao traduzir esses termos para o inglês, gera-se dúvida uma vez que os postos e graduações do USMC, por exemplo, são diferentes dos utilizados pela Marinha dos EUA (US Navy). Embora o USMC faça parte da Marinha dos EUA, seus postos correspondem aos do Exército dos EUA, salvas algumas exceções (NATO, 1996)[2]. A imagem a seguir corresponde a um extrato da equiparação entre os postos das forças armadas americanas (exceto a Força Aérea), conforme a padronização estabelecida pela Organização do Tratado do Atlântico Norte (OTAN):

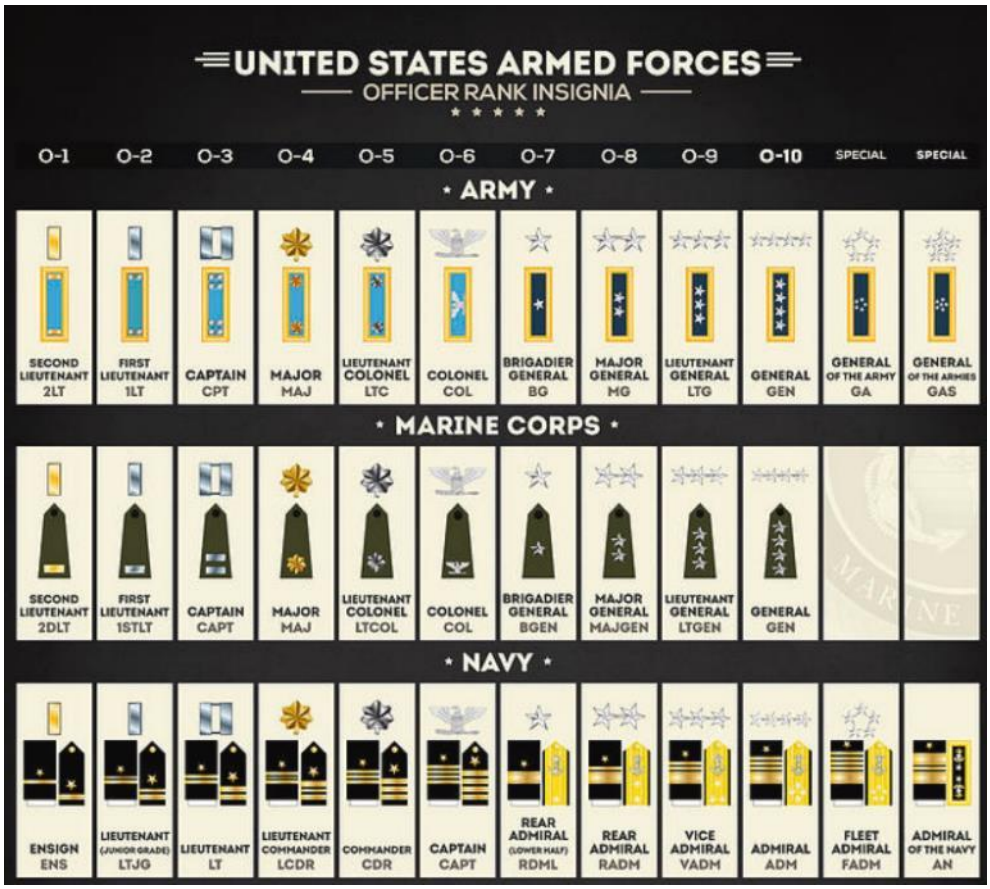


Figura 3 – Postos do Exército, da Marinha e do Corpo de Fuzileiros Navais dos EUA conforme padronização da OTAN. Fonte: MILITARY.COM, 2021a.

1. De acordo com BRASIL (2016), com o fim do período de transição para a implementação do Novo Acordo Ortográfico, deve-se grafar sem hífen os postos que contêm preposição, substituindo as grafias anteriores: Almirante-de-Esquadra, Capitão-de-Mar-e-Guerra, Capitão-de-Fragata e Capitão-de-Corveta.

2. A OTAN mantém uma lista padronizada de patentes militares a fim de tentar estabelecer uma correspondência entre a classificação militar de seus países membros. Essa lista de classificação padrão se encontra no documento STANAG 2116 (NATO, 1996), cujo nome formal é *NATO Codes for Grades of Military Personnel*.

Na figura 3, podemos confirmar tanto a semelhança entre a nomenclatura utilizada para os postos do Exército e do Corpo de Fuzileiros Navais americanos quanto a total diferença em relação à nomenclatura utilizada para os postos da Marinha dos EUA.

No Brasil, os produtos de terminologia militar bilíngue (inglês e português) disponíveis hoje ou não contêm a informação sobre os termos utilizados para postos estruturada e completa ou, ainda, a traz de maneira equivocada, embora estruturada. Esses produtos, da mesma forma, trazem informações contraditórias e desatualizadas.

O glossário *English-Portuguese Dictionary of Military Terminology (A-Z)* (U.S. ARMY, 2015) é uma obra de referência utilizada pelas Forças Armadas brasileiras para padronização de terminologia militar inglês-português. Esse material consiste em uma lista de palavras em inglês, dispostas em ordem alfabética, e seus equivalentes terminológicos para o português. Dentre essas palavras, há apenas alguns postos e graduações, conforme o exemplo mostrado na figura 4.

captain:	capitão (Army), capitão de mar-e-guerra (Navy), comandante (Merchant Marine)

Figura 4 – Verbete *captain*. Fonte: U.S. ARMY, 2015, p. 146.

O posto *captain* não aparece estruturado em uma tabela com as outras patentes e seus equivalentes em língua inglesa, de forma organizada, tendo indicadas suas respectivas forças, sejam elas do Brasil, EUA ou Reino Unido. Essa informação aparece em forma de verbete, em uma lista junto com outros termos organizados alfabeticamente, sem explicitar a relação hierárquica estabelecida com outros postos. Informa-se apenas, conforme exibido na figura 4, além da proposta de equivalente em língua portuguesa, que o termo em inglês é utilizado pela Marinha (que podemos inferir se tratar da Marinha americana, devido à obra ser dessa nacionalidade). Ainda, consta a grafia antiga (com hífen em *mar-e-guerra*) e sem letra maiúscula, conforme é grafado em português brasileiro.

Com isso, em relação aos postos, além dos problemas elucidados, restam algumas perguntas não respondidas pelo material, como as seguintes: uma vez que o CFN utiliza a mesma nomenclatura para postos e graduações que a MB, o que

não ocorre entre o USMC e a Marinha dos EUA, deve-se basear nos postos da Marinha americana ou do USMC para realizar a equivalência tradutória para o inglês? E, sabendo qual força de língua inglesa devo utilizar como referência, quais são os termos equivalentes que essa força utiliza para os postos? Qual seria o equivalente tradutório ao posto de *Segundo-Tenente*, por exemplo, que o material não traz?

Tratemos agora de outro material terminológico disponível no mercado, o *Dicionário de Expressões e Termos Militares* (MARINOTTO, 2011). Apesar de ser denominado *dicionário*, o produto não traz definições, mas equivalentes em inglês e português listados alfabeticamente, além de tabelas com informações adicionais, como acrônimos, abreviaturas, postos e graduações. A tabela a seguir apresenta um extrato do material, com os postos equivalentes entre as Marinha dos EUA, do Reino Unido e do Brasil.

ESTADOS UNIDOS [USM]	REINO UNIDO [RN]	BRASIL [MB]
Commissioned officers	Commissioned officers	Oficiais
Fleet Admiral	Admiral of the Fleet	Almirante [Alte]
Admiral [ADM]	Admiral [Adm]	Almirante-de-Esquadra [Atle Esq / AE]
Vice Admiral [VADM]	Vice-Admiral [V-Adm]	Vice-Almirante [V Alte / VA]
Rear Admiral [RADM]	Rear-Admiral [Rear-Adm]	Contra-Almirante [C Alte / CA]
Commodore [CDRE]	Commodore [Cdre]	Capitão-de-Mar-e-Guerra [CMG]
Captain [CAPT]	Captain [Capt]	Capitão-de-Fragata [CF]
Commander [CDR]	Commander [Cdr]	Capitão-de-Corveta [CC]
Lieutenant Commander [LCDR]	Lieutenat Commander [Lt-Cdr]	Capitão-Tenente [CT]
Lieutenant [LT]	Lieutenant [Lt]	Primeiro-Tenente [1 T]
Lieutenant Junior Grade [LTJG]	Sub-Lieutenant [Sub-Lt]	Segundo-Tenente [2 T]
		Guarda-Marinha [GM]

Quadro 1 – Equivalência entre os postos de Oficiais das Marinhas dos EUA, Reino Unido e Brasil, respectivamente. Fonte: Marinotto, 2011, seção: Hierarquia Militar.

O quadro acima, apesar de conter uma lista muito mais estruturada e organizada, evidencia algumas inconsistências. Nota-se que o posto *Commodore*, que aparece tanto na lista americana quanto na britânica, é equiparado ao posto da MB

de *Capitão-de-Mar-e-Guerra*[3][sic]. Porém, esse posto não conta na lista da OTAN (NATO, 1996) e, segundo consta no glossário do U.S. Army, 2015, o equivalente tradutório de *Capitão de Mar-e-Guerra*[4][sic]para a Marinha americana é *captain* (figura 4). Ainda, em Marinotto, 2011, *Capitão-de-Fragata*[sic] é equivalente a *Commander*; *Capitão-de-Corveta*[sic] é equivalente a *Lieutenant Commander* e *Capitão-Tenente* é equivalente a *Lieutenant* (p. 184), contradizendo novamente U.S. Army (2015). Ou seja, não há uma padronização na informação, além de, nos dois produtos, não constar a grafia atualizada quanto à mudança ortográfica de retirada do hífen dos termos preposicionados prevista em BRASIL (2016).

Outra inconsistência tradutória no âmbito do CFN ocorre ao se tentar traduzir o termo *praça*, conforme ocorre na obra bilíngue *Fuzileiros Navais: confie neles!*:

> No Centro de Instrução Almirante Sylvio de Camargo (CIASC) os Oficiais e as Praças do Corpo de Fuzileiros Navais (CFN) realizam, respectivamente, cursos de aperfeiçoamento (...)
>
> *All the Admiral Sylvio de Camargo Centre Officers and enlisted Marines go through their respective basic and advanced courses (...)* (MEDEIROS e OLIVEIRA, 2013, p.52, grifo nosso).

Uma vez que o ingresso no CFN brasileiro se dá apenas por concurso público (MEDEIROS, 2013; CHAGAS, 2021), e não apenas por alistamento, sem realizar testes de conhecimento, conforme ocorre no serviço militar obrigatório, o termo em inglês *enlisted*, conforme aplicado pelo tradutor na citação anterior, não é considerado, por esse motivo, o equivalente mais adequado por muitos militares. Assim, alguns preferem utilizar o termo britânico *other ranks* ou, ainda, o termo americano *Non-Comissioned Officers* (NCO). A segunda, porém, não pode ser utilizada para designar todas as praças, uma vez que se refere apenas a Sargentos e Cabos e não inclui Soldados (MILITARY.COM, 2021b). Essas duas últimas possibilidades tradutórias para *praça* foram utilizadas por Oficiais Fuzileiros Navais brasileiros durante a visita de uma comitiva da ONU ao Complexo Naval da Ilha do Governador, em 19 de julho de 2021.

Haveria, ainda, muitos outros exemplos a serem mencionados quanto a problemas tradutórios no âmbito do CFN, como em relação aos nomes de Organizações

3. Observa-se que o termo, segundo Marinotto, 2011, é escrito com 4 hifens.
4. Aqui, o mesmo termo, em U.S. Navy, 2015, p. 146, é escrito com 3 hifens.

Militares, ao uso de siglas e abreviações e às tradições navais. Portanto, diante de tantas questões, faz-se necessária a criação de um material terminológico não tão abrangente quanto os produtos já existentes voltados para o meio militar, mas direcionado às questões específicas do CFN. A fim de produzir um material que contenha não apenas termos possíveis, mas consagrados nesse campo de conhecimento, é preciso que seja realizado um trabalho cuidadoso com o auxílio e aprovação de especialistas da área, com equivalentes terminológicos extraídos de materiais de referência, e que possibilite padronizar a terminologia bilíngue português-inglês do CFN.

Assim, neste trabalho, propõe-se a criação de um léxico português-inglês de termos da subárea em estudo, o *Espírito de Corpus*. O título escolhido para o léxico remete, além do estudo com corpus, a um dos sentimentos mais importantes para os Fuzileiros Navais. Segundo BRASIL (2020b, grifo nosso), "[O CFN] cultiva com especial carinho o espírito de corpo, uma forma de pensar e uma crença que polarizam homens na busca de objetivos comuns".

Para a criação do léxico, realizaremos a apresentação de um método de extração e validação de terminologia bilíngue (português e inglês) específica de determinada subárea de domínio (o Corpo de Fuzileiros Navais do Brasil) com a utilização da Terminologia, dos Estudos da Tradução Baseados em Corpus, da Terminologia Computacional e de critérios de validação de termos baseados na Semântica Lexical.

Este estudo se divide em duas partes: estudo piloto e confecção do léxico. Primeiramente, foi realizado um estudo piloto com um corpus paralelo a fim de comparar o desempenho de duas ferramentas de Extração Automática de Termos: o AntConc 3.5.7 e o TermoStat Web 3.0. As ferramentas foram testadas com os subcorpora em português e em inglês da subárea do Corpo de Fuzileiros Navais. Na segunda etapa do estudo, foi realizada a extração e análise de termos de um corpus comparável para a confecção do léxico propriamente dito com o auxílio da ferramenta que se destacou nos testes e com o método que apresentou os melhores resultados.

Dessa forma, além de produzir material terminológico confiável e aprovado por especialistas da subárea de conhecimento militar em questão[5], buscou-se com esta pesquisa solucionar alguns problemas atribuídos à Extração Automática de Termos, assim como auxiliar no procedimento de validação manual de termos e na resolução de problemas tradutórios específicos do CFN.

Espera-se que este estudo venha a facilitar o trabalho de tradutores e especialistas que desejam realizar estudos similares para a investigação de corpora de áreas especializadas, de grande serventia para a posterior criação de material terminológico. Além disso, o produto gerado, o léxico *Espírito de Corpus*, é uma base de referência confiável, com 271 termos em português e seus equivalentes terminológicos em inglês validados por especialistas. Contém, ainda, informações essenciais de uso como notas, siglas/abreviações e variantes, e já pode ser consultado por tradutores e profissionais especializados na internet no endereço https://maritraduz.com/lexico/.

5. A autora desta dissertação é Oficial da Reserva Não Remunerada (RM2) do Quadro Técnico de Inglês da MB com experiência de oito anos servindo no CFN. Logo, é uma especialista da área. Além dela, revisaram o material o Capitão de Fragata (Fuzileiro Naval) Villarmosa, Oficial de Gabinete do Comando-Geral do CFN, e o Capitão de Fragata (Fuzileiro Naval) Maia, Oficial de intercâmbio do CFN no USMC de 2021 a 2022.

2
ASPECTOS TEÓRICO-METODOLÓGICOS

2.1
TERMINOLOGIA

A palavra Terminologia é polissêmica. Cabré (1999) apresenta a multiplicidade de significados da palavra ao explicar que a Terminologia engloba três acepções: a) conjunto de princípios e bases conceituais que regem o estudo dos termos (disciplina); b) conjunto de diretrizes utilizadas no trabalho metodológico (metodologia); c) conjunto de termos de uma área de especialidade (conjunto de termos em si). Ainda, para Krieger e Finatto (2004), a polissemia do termo é indicada na própria grafia da palavra: Terminologia é grafada com T maiúsculo ao se referir a um campo de estudos ou disciplina, enquanto a terminologia é grafada com t minúsculo ao se referir a um conjunto de termos. Neste trabalho, adotaremos essa diferenciação gráfica a fim de melhor contextualizar o termo.

Em se tratando de Terminologia (com t maiúsculo), diversos teóricos buscaram defini-la como estudo científico (LABATE, 2008; ANDRADE, 2001), campo teórico-prático (REY, 2007; KRIEGER e FINATTO, 2004; KRIEGER, 2001; LAFACE, 2001) ou ciência/disciplina científica (BARROS, 2004; MACIEL, 2001; WÜSTER, 1979, apud L'HOMME, 2020).

Dentre os teóricos que classificam a terminologia como estudo científico, destaca-se Labate (2008, p.15), que explica que "Terminologia é definida pela ISO 1087 como o estudo científico das noções e dos termos usados nas línguas de especialidade". Também para Andrade (2001, p.193): "A Terminologia é, antes de tudo, um estudo do conceito e dos sistemas conceptuais que descrevem cada matéria especializada".

Para os teóricos que enfatizam o campo teórico-prático da Terminologia, sem mencionar seu caráter científico, podemos citar Krieger e Finatto (2004, p.16), para quem:

> a Terminologia é um campo teórico-prático que estuda o conjunto de termos específicos de uma área científica e ou técnica, bem como direciona a produção de glossários, dicionários técnico-científicos e bancos de dados terminológicos.

Krieger (2001, p. 34), ainda, considera "a Terminologia como componente lexical das comunicações especializadas e expressão dos saberes técnicos e científicos".

Outros teóricos ressaltam o caráter prático da Terminologia. Para Rey (2007, p.330), a Terminologia é

> uma atividade baseada no reconhecimento de áreas organizadas do conhecimento, dividida ou distribuída em entidades semânticas delimitadas pelas definições e registradas em cada língua por meios essencialmente lexicais.

Esse caráter prático também é evidenciado por Laface (2001, p.240), para quem "[...] o contexto terminológico é antes de tudo uma prática que responde pelas necessidades sociais".

Já a criação da Terminologia como disciplina é atribuída a Wüster (1979, apud L'HOMME, 2020), mas os esforços para organizar o saber científico em estruturas e o desenvolvimento de métodos de padronização, conforme podemos observar na taxonomia de espécies animais em zoologia, por exemplo, vem ocorrendo muito antes dele. Seus princípios e reinterpretações de seus estudos fazem parte do escopo da Teoria Geral de Terminologia (*General Theory of Terminology* - GTT), comumente chamado de Terminologia. Wüster, porém, criou uma teoria que visa a assegurar uma comunicação sem ambiguidades. Para isso, duas suposições foram criadas: 1. De que o conhecimento tem uma estrutura. Logo, unidades linguísticas utilizadas para representar esse saber refletem essa estrutura; 2. De que comunicação do texto especializado se apoia em unidades linguísticas não ambíguas.

A Terminologia também é considerada uma ciência ou disciplina científica para Barros (2004, p.21): "a disciplina científica que estuda as chamadas línguas (ou linguagens) de especialidade e seu vocabulário." Já para Maciel (2001, p.39): "é a ciência que se ocupa do termo, unidade lexical 'profissionalmente marcada' [...] e se caracteriza por sua natureza inter e transdisciplinar".

Vários teóricos também comparam a Terminologia a outros campos de estudo, dos quais também pode usufruir, como a Lexicologia e a Semântica Lexical, segundo L'Homme (2020). Para a autora, por se tratarem de disciplinas próximas e se ocuparem do estudo das palavras, compartilham de várias semelhanças e oferecem contribuições umas às outras, uma vez que a Terminologia é o estudo dos termos, e uma vez que termos são tipos de palavras. Segundo ela:

> um número crescente de pesquisadores (Condamines 1993; Lerat 2002a; Gaudin 2003; Aldestein e Cabré 2002; Faber e L'Homme 2014; entre outros) enfatiza a utilidade da Semântica Lexical para a terminologia. Na verdade, a Semântica Lexical e a terminologia têm muito em comum, uma vez que ambas as disciplinas visam responder a perguntas sobre a natureza das palavras, o conteúdo da palavra (ou seja, a natureza do significado), a relação entre o conteúdo da palavra e nossa interpretação da realidade, e as relações entre as palavras e significados das palavras. (L'Homme, 2020, p.1, tradução nossa)[6]

Assim, neste trabalho, usaremos a abordagem de L'Homme ao unir a Semântica Lexical à Terminologia para a validação de candidatos a termos nos corpora de estudo. Porém, vale ressaltar a importante contribuição fornecidas por outras semânticas no âmbito do estudo dos termos, como a Semântica de *Frames*, teoria proposta por Fillmore no fim da década de 70 (FILLMORE, 1976).

2.1.1
TERMOS: OBJETO DE ESTUDO DA TERMINOLOGIA

Conforme exposto, em todas as suas definições e interpretações, a Terminologia tem como objeto de estudo (mas não somente) os termos que compõem as línguas de especialidade. Termos ou unidades terminológicas são palavras que dão nome a um conceito de determinada área de especialidade. Apesar de uma palavra de língua geral designar diversos significados de acordo com seu contexto, um termo designa um único conceito.

Na Terminologia, chamamos de conceitos os itens do conhecimento, que nada mais são do que generalizações de entidades extralinguísticas. Logo, os conceitos podem, teoricamente, ser considerados entidades que podem ser delineadas independentemente das denominações que recebem para representá-los.

Termos devem ser diferenciados de unidades que não possuem um status terminológico. Apesar de os terminólogos não serem normalmente especialistas do

6. No original: *an increasing number of researchers (Condamines 1993; Lerat 2002a; Gaudin 2003; Aldestein and Cabré 2002; Faber and L'Homme 2014; among others) stress the usefulness of lexical semantics for terminology. In fact, lexical semantics and terminology have much in common, since both disciplines aim to answer questions about the nature of words, word content (i.e. the nature of meaning), the relationship between word content and our construal of reality, and relations between words and word meanings.*

campo de conhecimento em que os termos em questão estão inseridos, os dois fatores de maior importância a serem considerados para a detecção dos termos são a área de conhecimento e a aplicação dos termos (L'Homme, 2020).

Embora haja um consenso maior em torno dos substantivos no que diz respeito à definição do status terminológico de um item lexical, os termos podem também caracterizar entidades expressas tipicamente com substantivos e adjetivos (classe de adjetivos relacionais, ex.: operação anfíbia); podem, ainda, denotar atividades, eventos ou processos que são expressos por verbos, e propriedades, expressas também por adjetivos e advérbios.

Portanto, segundo L'Homme (Ib. Idib.), os termos podem pertencer a uma dessas quatro classes gramaticais mencionadas: substantivo, verbo, adjetivo e advérbio. A autora afirma ainda que, para considerar um item lexical como termo, é necessário realizar uma análise minuciosa de seu significado, além de distinguir o seu significado de outros itens no caso de polissemia.

Quanto a sua composição, os termos podem ser simples, quando possuem apenas um radical ou unidade lexical (ex.: comandante); compostos, quando possuem dois ou mais radicais (ex.: assalto anfíbio); e complexos, quando possuem dois ou mais radicais, além de outros elementos (ex.: componente de combate terrestre). Segundo Krieger e Finatto (2004), os termos complexos têm maior ocorrência em um corpus (cerca de 70% do texto) do que os termos simples. Dentre os termos complexos, sua maioria é formada por sintagmas nominais.

Quanto ao tamanho das expressões regulares (ou n-gramas) de um corpus, podemos chamar itens lexicais de tamanho um (ou seja, que contêm um só elemento) de unigramas; de tamanho dois, de bigramas; de tamanho três, de trigramas; e assim por diante (TEIXEIRA, 2010).

No âmbito dos termos compostos, podemos identificar as colocações, que nada mais são do que palavras que possuem entre si uma ligação semântica devido a sua coocorrência frequente (BUSSMANN, 2006). Quando se combinam, essas palavras, que se limitam a duas unidades, estabelecem entre si uma relação de base (palavra de maior carga semântica) e colocado (palavra que acompanha a base), segundo Hausmann (1985, apud TAGNIN, 2002). Ou seja, as colocações representam combinações recorrentes de palavras, como *fuzileiro naval* e *operação anfíbia*, aprendidas naturalmente pelos falantes de determinada língua (TAGNIN, 2002).

Quanto à organização dos termos, uma vez que são considerados rótulos linguísticos, eles podem ser ordenados nas chamadas estruturas conceituais, mapas conceituais ou árvores de domínio. Tais estruturas são construídas por meio do consenso entre especialistas da área em questão ou terminólogos, os quais, muitas vezes, precisam criar categorias para organizar esse conhecimento.

Uma árvore de domínio, segundo Barros (2004, p. 3224), "é um diagrama hierárquico composto por termos-chave de uma especialidade, semelhante a um organograma". Sua utilização é recomendada pelas normas ISO referentes a trabalho terminográfico (ISO 1087, 2000, e ISO 5127, 2017) a fim de possibilitar uma aproximação primária da área de conhecimento com a qual se vai trabalhar. Esse mapa conceitual, ainda, visa apenas a oferecer uma organização possível para determinada especialidade, para que o pesquisador se baseie nele a fim de compreender algumas das hierarquias básicas desse campo de estudo, além de situar o recorte em que atuará. Assim, a estruturação dessa área de conhecimento, identificando suas subáreas e interrelações, é ponto de partida para a geração de listas de termos, ao condicionar o seu reconhecimento. Essa estruturação precisa ser feita a partir de dados coerentes, atualizados e cumprir normas de qualidade (PAVEL e NOLET, 2002).

Usaremos como ponto de partida, neste estudo, após a seleção dos corpora de estudo, nossa árvore de domínio do CFN constante na monografia *Estudo terminológico bilíngue de termos militares da subárea do Corpo de Fuzileiros Navais* (MÜLLER, 2019), aprovada por especialistas da área. Essa árvore está representada no gráfico a seguir, indicando as subáreas abrangidas em cinza e os termos já inseridos em verde no estudo mencionado.

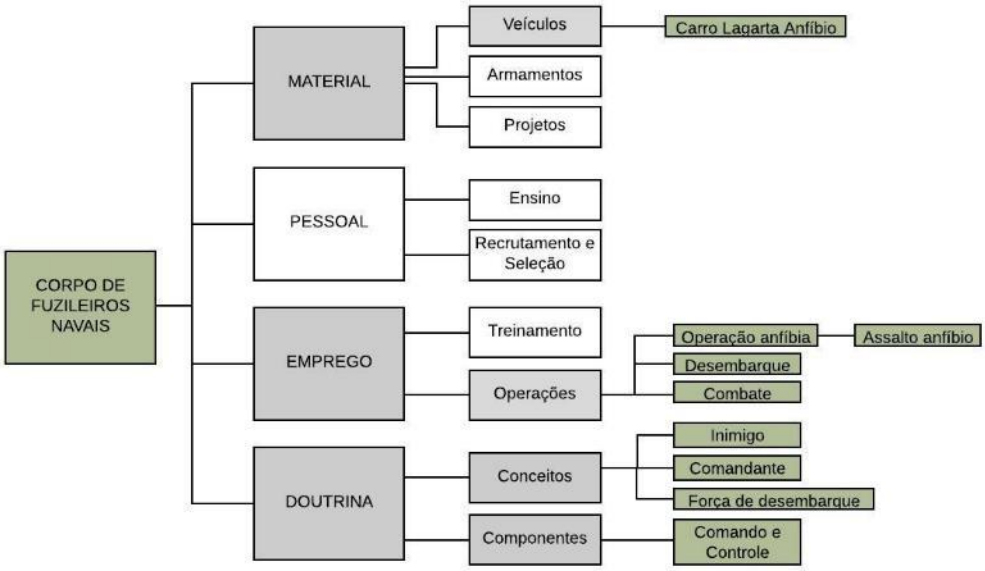


Figura 5 – Árvore de domínio do CFN preenchida com termos. Fonte: Müller, 2019.

A árvore da figura 5 foi modificada neste trabalho devido às diferenças entre os estudos. Enquanto em Müller (2019) o propósito foi de realizar o estudo de apenas dez termos do CFN por meio da criação de fichas terminológicas bilíngues, aqui, o objetivo foi de criar um léxico bilíngue muito mais extenso, que ao final apresentou 271 termos com seus respectivos equivalentes terminológicos em inglês, além de informações adicionais. A árvore de domínio atualizada pode ser vista no capítulo 5 desta pesquisa.

2.1.2
EQUIVALÊNCIA EM TERMINOLOGIA

A equivalência é um conceito bastante controverso e polêmico utilizado por muitos teóricos da tradução, como Catford (1965), Nida (1993) e Toury (1995). Gideon Toury caracteriza o conceito de equivalência como um processo sociocultural. O autor propõe analisar se a tradução, na relação com o texto original, é direcionada para a língua de partida ou para a língua de chegada. Já para Pym, aquilo que podemos dizer em uma língua "pode" apresentar o mesmo valor quando traduzido para outra língua:

> A relação entre o texto de partida e a sua tradução será de equivalência (igual valor), onde valor pode estar no nível da forma, da função ou em algum intermediário entre elas. (2017, p. 27).

Para falar de equivalência, conforme aponta Adamska-Sałaciak (2010, p. 387, apud PIMENTEL, 2013) é preciso haver (no mínimo) duas entidades de algum tipo, uma certa relação entre essas entidades e um certo valor nessa relação. Pimentel (2013) aponta a contribuição de diversos autores em relação ao estudo da equivalência para a Terminologia (Adamska-Sałaciak, 2010; Atkins e Rundell, 2008; Hartmann, 2007 [1990]; Piotrowski, 1994; Svensén, 2009; Werner, 1999; Wiegand, 2005, apud Pimentel, 2013), os quais buscaram responder as muitas questões existentes em torno dessas entidades e as relações entre elas.

Ainda, a natureza de equivalentes terminológicos normalmente não é tratada pela Terminologia devido ao conflito gerado pela relação entre a equivalência e o entendimento da relação unívoca de um conceito para um termo. Nesse âmbito, Sager (1994, apud L'HOMME, 2020, p. 229) explica que:

> Na terminologia, encontramos pouca ou nenhuma referência à natureza dos equivalentes de tradução, porque a teoria da referência monossêmica de termos a conceitos não admite prontamente problemas de equivalência. [7](tradução nossa)

Convém ressaltar que este trabalho não pretende se aprofundar na questão da equivalência, mas mencionará apenas alguns aspectos da abordagem proposta por L'Homme (2020) para tratar da questão, conforme sugere a Semântica Lexical – metodologia escolhida para validação dos termos neste trabalho, conforme justificado mais à frente. Nesse âmbito, a autora menciona duas abordagens para se tratar a equivalência em Terminologia: a abordagem baseada no conhecimento (equivalência conceitual) e a abordagem baseada no léxico (equivalência terminológica), embora reconheça que não há uma divisão clara entre as duas.

A abordagem baseada no conhecimento diz respeito à relação entre termos que pertencem a línguas diferentes e denotam o mesmo conceito dentro do mesmo domínio. Assim, segundo a autora, os equivalentes devem ser lançados no registro do mesmo termo em bancos de termo ou em uma mesma ficha terminológica. Tal

7. No original: *In terminology we find little, if any reference to the nature of translation equivalents, because the theory of monosemic reference of term to concepts does not readily admit to problems of equivalence.*

abordagem preocupa-se com uma equivalência exata, considerando uma relação em que termos de diferentes línguas denotam um único conceito.

Já a abordagem baseada no léxico estabelece a relação entre termos de diferentes línguas e que transmitem o mesmo significado dentro do mesmo domínio. O problema é que os itens lexicais podem ser polissêmicos, o que afeta a equivalência, daí a relação conflituosa entre a Terminologia e a equivalência, mencionada anteriormente.

Por essa razão, em abordagens baseadas no léxico, a equivalência exata é definida entre duas unidades lexicais (itens lexicais com um significado específico), não entre dois itens lexicais. O resultado dessa abordagem é uma rede muito mais complexa de equivalência.

Logo, segundo a autora, três situações diferentes podem ocorrer ao considerar itens lexicais em duas línguas diferentes. Primeira situação: um item lexical carrega (pelo menos) dois significados diferentes, estando o primeiro associado ao léxico de língua geral, e, o segundo, a um domínio especializado. Por exemplo, observemos o item lexical *key*. Na língua geral, consiste em um objeto de metal utilizado para abrir portas. Na computação, consiste em uma parte de um teclado pressionado por um usuário para inserir um caractere ou enviar um comando. Segunda situação: um item lexical transmite (pelo menos) dois significados diferentes ligados a diferentes áreas do conhecimento. Exemplo: transmissão (telecomunicações X virologia). Terceira situação: um item lexical carrega (pelo menos) dois significados diferentes que coexistem no mesmo domínio. Exemplo: comando (no domínio militar: 1. Ordem, instrução ou 2. Conjunto de instâncias militares superiores).

Estudos recentes (L'HOMME, 2004; 2020) mostram que as abordagens léxico-semânticas são eficazes para se trabalhar com dados extraídos de corpora por se basearem na observação das relações entre os itens lexicais conforme aparecem no contexto de origem; logo, procura-se estabelecer uma relação de equivalência entre significados em um determinado domínio específico. Neste estudo, por tratarmos da equivalência terminológica, buscaremos aplicar a abordagem baseada no léxico ao tratarmos de significados específicos dentro do contexto do CFN. Eventuais casos de polissemia encontrados entre os idiomas dentro do mesmo domínio, conforme descritos em 5.4, serão lançados na mesma entrada do léxico.

2.1.3
PROBLEMAS DE EQUIVALÊNCIA ENTRE IDIOMAS

Segundo a ISO 10241, *termos equivalentes* são termos em línguas diferentes que representam o mesmo conceito. Para o registro em recursos terminológicos (dicionários, glossários, bancos de termos etc.), busca-se a equivalência exata. No entanto, diferentes situações podem impedir que sejam encontrados termos equivalentes. Vamos mencionar três possibilidades: princípio da não equivalência, princípio da equivalência parcial e diferenças estruturais entre idiomas (L'HOMME, 2020).

A não equivalência ocorre quando a língua B não possui equivalente adequado para expressar o significado transmitido por um termo na língua A. Por exemplo: o termo *hashtag*, em inglês, no contexto das redes sociais, que não possui equivalente em português. A solução encontrada no português foi de realizar o empréstimo direto do termo em inglês.

Culturas diferentes, ao longo do tempo, acabam ajustando, adaptando ou criando equivalentes em sua própria língua para tais termos por meio de diferentes estratégias, que podem ser um empréstimo direto (como em *hashtag*) com ou sem uma explicação na língua meta, sua tradução literal, uma paráfrase, uma adaptação ou criação de uma nova designação (L'HOMME, Id. Ibid., p. 235).

Já a equivalência parcial ocorre quando uma língua faz uma distinção que não é feita na outra. Exemplo clássico: *wood*. Em português: lenha (madeira usada para aquecimento) e madeira (material utilizado para construção). Em inglês, há apenas a palavra *wood* para os dois significados. Porém, há equivalência perfeita entre os idiomas português e espanhol (*leña*, *madera*) (L'HOMME, Id. Ibid., p. 236).

A autora aponta também que nem sempre os equivalentes pertencem à mesma parte do discurso. As divergências estruturais ocorrem quando um significado expresso por um termo que pertence a uma determinada parte do discurso não possui equivalente direto em outro idioma (diferenças de equivalências gramaticais, sintáticas etc.). Exemplo: *Ele acabou de chegar*. / *He has just arrived* (*acabou de* – equivalente a *just*).

2.1.4
TERMINOGRAFIA E PRODUTOS TERMINOLÓGICOS PARA PROFISSIONAIS DE LÍNGUAS

Ao se ocupar do estudo dos termos e de como eles se comportam nos textos, a Terminologia segue determinados procedimentos metodológicos. Cabré (1999, p. 10) lista e descreve a importância desses procedimentos, que envolvem os processos de compilação, descrição, processamento e apresentação dos termos envolvidos no trabalho terminológico:

> Definida como o processo de compilar, descrever, processar e apresentar os termos de campos temáticos especiais em uma ou mais línguas, a terminologia não se esgota em si mesma, mas atende às necessidades sociais e tenta otimizar a comunicação entre especialistas e profissionais, assistindo diretamente a tradutores ou a outros profissionais preocupados com a padronização de um idioma. (tradução nossa)[8]

Dessa forma, ressalta a colaboração que a Terminologia fornece aos profissionais que trabalham com línguas e dependem da disciplina para facilitar o processo de comunicação ao qual promovem a intermediação. Tais profissionais usuários de terminologias são divididos em dois grupos principais: o de usuários diretos e o de profissionais de línguas.

Os usuários diretos, que são os especialistas das áreas de conhecimento, necessitam diretamente da padronização dos termos para nomear e definir os conceitos com que trabalham, a fim de favorecer a comunicação em seus campos de atuação. Já para os profissionais de línguas, como tradutores, escritores e intérpretes, as terminologias permitem facilitar o processo de comunicação com que atuam. Para eles, léxicos e vocabulários bilíngues e multilíngues, glossários e dicionários especializados são ferramentas essenciais que permitem aperfeiçoar a sua atividade profissional.

Gomez e Vargas (2004, apud TAGNIN, 2009, p. 1081) ressaltam que "materiais terminológicos deveriam auxiliar o tradutor nas tomadas de decisão a que

8. No original: *Defined as the process of compiling, describing, processing and presenting the terms of special subject fields in one or more languages, terminology is not an end in itself, but addresses social needs and attempts to optimize communication among specialists and professionals by providing assistance either directly or to translators or to committees concerned with the standardization of a language.*

está sujeito em sua prática diária". Porém, tais ferramentas de apoio nem sempre estão disponíveis no mercado, ou se estão, nem sempre suprem necessidades específicas dependendo do trabalho e da área de atuação. Por isso, é comum que tradutores precisem atuar também como terminológo, montando o próprio material terminológico de forma rápida para aprimorar a sua atuação, conforme evidenciado, ainda, por Cabré (1999, p. 48):

> Para fazer seu trabalho, os tradutores dependem de vocabulários bilíngues ou multilíngues dos termos que ocorrem no texto. Isso não significa, no entanto, que os tradutores não preparem a terminologia por conta própria. Ocasionalmente, eles têm que agir como terminológo para encontrar equivalentes para os termos que não estão listados nos vocabulários disponíveis nem em bancos de dados especializados. Além disso, as restrições de tempo com as quais os tradutores muitas vezes têm de trabalhar podem não permitir que entreguem a tarefa a um terminológo. (tradução nossa)[9]

A aplicação prática da Terminologia para a criação de material terminológico é chamada de Terminografia. É a ciência que se ocupa da criação de dicionários, glossários, léxicos e bancos de dados, por exemplo. Diferentemente da Lexicografia, que tem como finalidade a compilação de palavras em geral, podendo essas incluírem termos, a Terminografia se ocupa apenas das palavras de cunho especializado.

Os preceitos teóricos tanto da Lexicografia quanto da Terminologia oferecem orientação quanto à arquitetura de um produto terminológico no que diz respeito a sua macro e microestrutura. A macroestrutura se refere à forma como as entradas desse léxico são organizadas, a quantidade de termos que comporão as entradas, além das partes complementares, como introdução e anexos como imagens, organogramas etc. Para Barros (2004, p. 151):

> Por macroestrutura entende-se a organização interna de uma obra lexicográfica ou terminográfica. Esse tipo de organização está relacionado às características gerais do repertório, ou seja, à estruturação das informações em verbetes (que podem se suceder vertical e/ou horizontalmente), à presença ou não de anexos, índices remissivos, ilustrações, setores temáticos, mapa conceptual e outros.

9. No original: *To do their job translators depend on bilingual or multilingual vocabularies of the terms occurring in the text. This does not mean, however, that translators do not prepare terminology themselves. On occasion they have to act as terminologists to find equivalents for those terms that are not listed in the available vocabularies nor in specialized data banks. Besides, the time constraints within which translators often have to work may not allow them to hand the task over to a terminologist.*

Barros (Ib. Idib.) e Frübel (2006) indicam que a ordem tradicional de organização dos verbetes é a alfabética. Krieger e Finatto (2004) apontam ainda que, além da ordem alfabética, a organização dos verbetes pode ser feita por área temática e subtemática, de acordo com a estrutura conceitual de um domínio, conforme análise de especialistas da área do conhecimento abrangida no material.

Já a microestrutura se refere às informações contidas nos verbetes e a sua organização, como o próprio lema e o conjunto de informações sobre ele, os equivalentes em outro idioma, sua definição, exemplos reais de uso, suas variações terminológicas e possíveis relações semânticas que estabeleçam com outros itens lexicais. Segundo Barros (2004, p. 156):

> [...] a microestrutura compreende a organização dos dados contidos no verbete, ou melhor, o programa de informações sobre a entrada disposto no verbete. Três elementos devem ser levados em consideração, quando da distribuição dos dados na microestrutura: a) o número de informações transmitidas pelo enunciado lexicográfico/terminográfico; b) a constância do programa de informações em todos os verbetes dentro de uma mesma obra; c) a ordem de sequência dessas informações.

Muitas áreas e subáreas de conhecimento, como o Corpo de Fuzileiros Navais, subárea militar, possuem pouco material especializado bilíngue disponível sobre sua terminologia, o que gera dúvidas quanto à utilização e à padronização do vocabulário especializado a respeito, tornando necessária a criação de material terminológico para as áreas de especialidade envolvidas. Neste estudo, a fim de suprir a necessidade de criação de um compilado de termos bilíngue sobre o tema em foco contendo equivalentes em duas línguas, sem conter definições, optou-se por criar um léxico e não um vocabulário ou dicionário. Cabré (1989) explica a diferença entre os três produtos terminológicos: léxico trata de conjuntos de termos sem definições e que possui equivalentes em uma ou mais línguas. Quando esse conjunto de termos contém definições, consiste então em um vocabulário ou dicionário. Quanto mais exaustivo de informações, mais o material se aproxima de um dicionário e se afasta de um vocabulário.

O léxico bilíngue fruto deste trabalho se fundamenta na definição de Boutin-Quesnel et al. (1985, p. 30), que descreve um léxico como “um repertório que registra termos acompanhados de seus equivalentes em uma ou mais línguas, e que não apresenta definições. Nota: os léxicos, em geral, abrangem um só domínio”.

Sua estrutura, com base na explanação de Barros (2004), foi feita da seguinte forma: quanto à macroestrutura, foi feita a ordenação dos 270 verbetes que compõem o léxico por ordem alfabética, contendo também um texto introdutório explicando brevemente o que é o léxico, além de uma tabela e um organograma como anexos. Quanto à microestrutura, o léxico possui, além dos lemas, as possíveis variações terminológicas, como siglas e abreviaturas, e os equivalentes em inglês, acompanhados também de suas variações nesse idioma.

2.2
LINGUÍSTICA DE CORPUS, TRADUÇÃO E TERMINOLOGIA

Para Sinclair, referência na área de Linguística de Corpus e criador do primeiro dicionário realizado com o auxílio de um corpus computadorizado, o COBUILD, um corpus é definido da seguinte forma:

> (...) é uma coletânea de extratos de textos em formato eletrônico, selecionados de acordo com critérios externos para representar, na medida do possível, um idioma ou uma variedade de idiomas como fonte de dados para pesquisas linguísticas (2005, p. 1).[10]

Logo, corpora são grandes conjuntos de textos em linguagem natural, compilados a fim de atender a determinados objetivos e finalidades linguísticas. Já a Linguística de Corpus, para McEnery e Hardy (2012), é o estudo desses dados linguísticos em grande escala por meio de análise realizada com auxílio do computador.

Ao se unir à Terminologia, a aplicação da Linguística de Corpus é de grande utilidade para a tradução, uma vez que, por meio da utilização de ferramentas eletrônicas, permite processar grandes quantidades de dados, evidenciar como se comporta um termo em contextos naturais de linguagem, identificar e classificar itens quanto a diversos parâmetros etc. (BIBER et al., 1998; STUBBS, 1996). Dessa forma, são também fonte de grande utilidade para terminólogos para o estudo dos termos. Para complementar as informações extraídas desses textos, é necessário o auxílio de outras fontes, como especialistas do campo de conhecimento em estudo.

10. No original: *A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.*

2.2.1
TIPOS DE ESTUDO COM CORPORA

Há dois tipos de estudos envolvendo corpus: estudo baseado em corpus (*corpus-based*) e estudo dirigido por corpus (*corpus-driven*) (TOGNINI-BONELLI, 2001). Os estudos baseados em corpus geralmente utilizam dados de corpora a fim de explorar uma teoria ou hipótese já existente com o intuito de validá-la, refutá-la ou, ainda, refiná-la. Já os estudos guiados por corpus rejeitam a caracterização da Linguística de Corpus como um método e assegura, em vez disso, que o próprio corpus deve ser a única fonte de hipóteses sobre a linguagem. Afirma-se, portanto, que o corpus em si contenha sua própria teoria da linguagem. No entanto, McEnery e Hardy (2012) refusam tanto essa controvérsia gerada pela distinção binária entre os estudos com corpora quanto à perspectiva de que um corpus possua por si só um estatuto teórico, considerando, assim, todos os estudos com corpora linguísticos como baseados em corpus.

Logo, segundo a visão de McEnery e Hardy (Id. Ibid.), quando a Linguística de Corpus é utilizada como metodologia para a produção de um produto terminológico, dizemos que ela é baseada em corpus. Por meio dessa metodologia, a construção dos corpora é realizada de maneira minuciosa a fim de coletar textos representativos e que auxiliem no cumprimento do objetivo do pesquisador da melhor forma possível, com um material rico em informações e definições.

Quanto aos idiomas de sua composição, os corpora podem ser monolíngues ou não-monolíngues. De acordo com Dayrell (2005) os corpora não-monolíngues podem ser (1) multilíngues, (2) paralelos ou (3) comparáveis. Ainda, um corpus multilíngue contém textos produzidos em seus respectivos idiomas, cuja seleção segue as mesmas especificações, de forma que os textos possuam características semelhantes e possam ser contrastados. Tal contraste permite acesso a padrões linguísticos e ao consequente reconhecimento de estruturas equivalentes presentes nos idiomas de estudo. Esse material favorece os tradutores ao permitir, assim, a extração de terminologia, no caso dos textos especializados.

Da mesma forma, McEnery e Hardy (2012) afirmam que um corpus comparável pode ser definido como

> um corpus cujos textos foram coletados por meio da utilização do mesmo método de amostragem, por exemplo, as mesmas proporções dos textos dos mesmos gêneros nos mesmos domínios em uma gama de diferentes línguas no mesmo período de amostragem (p.20). [11]

Assim, ainda que esse conjunto de corpora possa ser definido como um corpus multilíngue, suas subpartes também podem ser consideradas corpora monolíngues. Caso haja uma equivalência de quadros de amostragem entre corpora em diferentes idiomas, eles podem ser considerados como monolíngues ou multilíngues, conforme necessário. Os subcorpora de um corpus comparável não consistem em traduções um do outro, mas sua comparabilidade está na semelhança entre seus critérios de amostragem.

Já um corpus paralelo, ainda segundo McEnery e Hardy (Ib. Idib.), é aquele que contém textos de origem em um idioma nativo (L1) e suas traduções (L2), assumindo que corpora paralelos são unidirecionais (por exemplo, do português para o inglês).

Quanto aos corpora de referência, Tagnin (2010, p. 358) define-os como um "corpus que serve de termo de comparação para o corpus de estudo. Em geral, deve ter três a cinco vezes o tamanho do corpus de estudo". Baker (2010) revela, ainda, que quanto maior o corpus de referência, mais palavras-chave, ou candidatos a termos, serão identificados. Quanto à especificidade desse corpus de referência, Scott (1999, apud BAKER, 2010), revela que, independentemente da área de conhecimento desse corpus, seja ele composto por romances ou textos médicos, por exemplo, a lista de palavras-chave encontradas serão relativamente semelhantes.

2.2.2
CARACTERÍSTICAS PRIMORDIAIS DE UM CORPUS

A tabela a seguir agrupa de forma resumida as características primordiais que um corpus deve possuir, fruto de um consenso crescente entre pesquisadores, segundo McEnery, Xiao e Tono (2006):

11. No original: (...) *a corpus containing components that are collected using the same sampling method, e.g. the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period.*

Origem	Dados autênticos
Propósito	Estudo linguístico
Composição	Conteúdo criteriosamente escolhido
Formação	Dados legíveis digitalmente
Representatividade	Deve ser representativo para determinada língua

Quadro 2 - Características indispensáveis a um corpus.

Quanto à representatividade do corpus em relação à sua extensão, Sinclair (1991) afirma que o corpus deve ser o maior possível dentro do que a tecnologia da época permitir. Em contrapartida, Sardinha ressalta que, quando um estudo tem como foco as obras de um único autor, ou, ainda, de uma única subárea dentro de uma área de conhecimento maior (como o CFN está para a área militar), caracteriza-se esse corpus, ainda que pequeno, como representativo, uma vez que "os corpora compilados em pequena escala por pesquisadores individuais acabam sendo mais representativos do que os respectivos subcorpora dos corpora gerais" (SARDINHA, 2000, p. 348). Mais importante do que o tamanho do corpus é quão bem projetado e representativo ele é. Logo, não há tamanho ideal para um corpus; tudo depende do seu conteúdo e do que está sendo investigado (FLOWERDEW, 1998, apud O'KEEFFE e MCCARTHY, 2012).

Por isso, corpora especializados são geralmente direcionados e configurados de forma cuidadosa para refletir recursos contextuais que atendam aos objetivos do pesquisador. Dessa forma, a análise de tais corpora pode revelar conexões entre padrões linguísticos e contextos de uso.

2.2.3
BENEFÍCIOS DOS CORPORA BILÍNGUES PARA A TRADUÇÃO

Ao se trabalhar com corpora bilíngues, a utilização da Linguística de Corpus permite encontrar equivalentes terminológicos nos idiomas de trabalho. Tognini-Bonelli (2001) destaca a importância da investigação das linhas de concordância de um corpus para a extração de termos. Como linhas de concordâncias, entendem-se aqui listagens de um item lexical e o seu cotexto, ou o texto que o rodeia, organizadas em linhas, nas quais a palavra de busca fica em evidência.

Tognini-Bonelli (2001) afirma que, uma vez que as palavras não ocorrem isoladamente, mas estabelecem relações de significado e estrutura com as palavras

ao seu redor, é possível identificar padrões no contexto em que estão inseridas. Dessa forma, os termos não podem ser identificados sozinhos, mas por meio da identificação das unidades multipalavras que se formam por meio dos padrões lexicais, gramaticais e o campo semântico no contexto em que se inserem. Assim, somente quando essas unidades completas são identificadas, é possível estabelecer relações de equivalência tradutória ou de comparação de unidades de significado entre as línguas de estudo.

Kübler e Aston (2010) e Kenning (2010) apontam dois tipos de corpora úteis para a pesquisa em tradução: o corpus paralelo, contendo textos originais na língua 1 (L1) e suas respectivas traduções na língua 2 (L2); e o corpus comparável, contendo textos semelhantes, em contexto original, em seus respectivos idiomas de estudo, mas que não são traduções.

O uso de corpora é capaz de auxiliar o tradutor durante todo o processo tradutório, segundo Kübler e Aston (Ib. Idib.). Os autores reforçam a importância da consulta aos corpora para esse fim, pois, ao contrário de dicionários, especialistas ou da própria internet, os corpora podem fornecem dados que não são previamente interpretados ou manipulados, mas consistem em amostras reais de texto, as quais permitem que os tradutores se familiarizem não só com os termos e conceitos do domínio, no caso do texto especializado, mas também com as convenções daquele gênero textual.

Tognini-Bonelli (2001) ressalta a importância da utilização de corpora comparáveis para uma identificação mais aprimorada e completa dos termos conforme utilizados em seus contextos. A autora propõe as seguintes etapas na análise de corpora bilíngues: primeiramente, deve-se identificar uma palavra ou expressão e as relações que estabelece em seu contexto. Após, deve-se comparar os corpora nas duas línguas de estudo. Nessa etapa, sugere-se, quando disponível, a utilização de um corpus paralelo para melhor identificação dos termos e seus equivalentes terminológicos. Os corpora comparáveis permitem, a partir daí, analisar os termos identificados dentro de seu contexto original, possibilitando analisar os equivalentes terminológicos identificados na etapa anterior e verificar se são adequados.

Em relação à comparação entre os corpora em busca de equivalentes terminológicos, Tognini-Bonelli (Ib. Idib.) afirma que o corpus evidenciará, nas linhas

de concordância, se os termos considerados tradução *prima-face* (literal) são realmente equivalentes terminológicos. Nos casos em que não houver uma tradução prima-face, a busca se dará pelos colocados ou pelo contexto.

Assim, ao analisar as palavras que acompanham os termos e as relações que estabelecem com o texto que acompanha os itens lexicais nas linhas de concordância do corpus, pode ser identificado qual é o melhor equivalente tradutório, evidenciando, por exemplo, os adjetivos e advérbios que co-ocorrem com mais frequência com o substantivo que se pretende traduzir.

A autora esclarece ainda que, se mesmo após a comparação das palavras-chave nas duas línguas de estudo, não for encontrado um equivalente satisfatório, por vezes até pela inexistência de um na língua meta, pode-se sugerir uma adaptação ou, ainda, inserir essa informação em uma nota.

É relevante mencionar que a extração de termos e a busca por equivalentes terminológicos por meio da combinação do uso de corpora paralelo e comparável já foi realizada com sucesso no estudo de Paiva, P. T. P., Camargo, D. C. e Xatara, C. M., 2008, que teve como objetivo criar um léxico bilíngue na subárea de cardiologia.

A seguir, veremos mais profundamente o que é e como é feita a extração automática de termos para se chegar a essas listas de palavras-chave por meio dessa metodologia.

2.3
EXTRAÇÃO MANUAL E AUTOMÁTICA DE TERMOS

Segundo a maioria dos pesquisadores, os termos podem ser analisados de um ponto de vista qualitativo e quantitativo, apresentando características distintas. A partir do ponto de vista qualitativo (BECˇKA, 1972; BAKER, 1988; FARRELL, 1990; NATION, 2001 apud CHUNG, 2003), é possível detectar termos a partir das características em comum entre eles, como as seguintes:

- muitos termos têm origem no latim e no grego;
- não serem utilizados em língua geral;
- sua particularidade em relação a determinado domínio;

- a polissemia que ocorre quanto ao seu significado em relação à língua geral, como *meios* (força e elementos materiais que integram o poder de combate)[12] e *meios* (bens, fortuna)[13].

Já do ponto de vista quantitativo (YANG, 1986; BAKER, 1988; FARRELL, 1990; GAMPER e STOCK, 1998/1999; KAGEURA e UMINO, 1996; NATION 2001 apud CHUNG, 2003), podem-se detectar os termos pelos seguintes fatores:

- a frequência do uso dos termos dentro de um assunto específico;

- a maior frequência de utilização dos termos em um domínio de especialidade do que em língua geral;

- a possibilidade de haver uma maior frequência dentro de um texto relacionado a determinada área do conhecimento;

- a concentração de sua frequência dentro de determinado tópico de um texto.

Observadas essas características qualitativas e quantitativas, é possível detectar graus “e "tecnicidade" dos termos em análise, dependendo de quão íntimo esses termos estão relacionados a um determinado domínio (NATION, 2001 apud CHUNG, 2003).

Dessa forma, utilizando os dois pontos de vista mencionados, podemos distinguir termos de não-termos ao avaliar o significado da palavra e ao analisar a gama e a frequência das formas de palavras existentes, comparando-as entre um corpus de um domínio de especialidade e um corpus de referência, para a partir daí realizar sua extração.

A detecção e a extração de termos são procedimentos que podem ser realizados de forma manual ou automática. Para Almeida, Aluísio e Oliveira (2007), a coleta (ou extração) de termos é a primeira etapa do método estabelecido para a realização de qualquer trabalho terminológico cuja finalidade seja produzir material terminográfico. Essa etapa é descrita da seguinte forma:

> a **extração de termos** diz respeito à obtenção do conjunto terminológico que comporá a nomenclatura do glossário ou dicionário. As fontes a partir das quais serão extraídos os termos devem ser previamente selecionadas, preferencialmente, devem ser fontes indicadas pelos próprios especialistas da área-objeto. A extração

12. Definição disponível em BRASIL, 2015.
13. Definição disponível em MEIOS, 2021.

> pode ser feita de forma manual ou automática, entretanto, quando se utiliza extração automática, é necessária a elaboração de corpus em formato digital, evidentemente. (ALMEIDA, ALUÍSIO E OLIVEIRA, 2007, p. 410, grifo dos autores).

A Terminologia e a Linguística de Corpus têm se desenvolvido de forma a aprimorar e automatizar muitos dos procedimentos metodológicos que utiliza, especialmente quanto à extração de termos. Graças aos recursos computacionais, cada vez mais evoluídos, hoje é possível manipular uma grande quantidade de dados de forma cada vez mais precisa, algo impossível por meio da extração manual. Diante da crescente facilidade de se adquirir textos em formato eletrônico na Web, tem sido frequente a utilização de corpora nas pesquisas terminológicas como objeto de extração automática de candidatos a termos (EAT).

A EAT visa a extrair unidades terminológicas de corpora considerando dados quantitativos. O resultado das extrações dos termos encontrados em um corpus gera matéria prima de grande valia para a elaboração de material linguístico e computacional. Para tal, podemos nomear três técnicas de EAT: linguística, estatística e híbrida.

Os sistemas baseados em conhecimento linguístico (HEID et al., 1996; KLAVANS e MURESAN, 2000, 2001a, 2001b apud ALMEIDA, ALUÍSIO e OLIVEIRA, 2007) utilizam recursos que contam com diferentes informações linguísticas para a extração dos termos, como informações lexicográficas, morfológicas, morfossintáticas, semânticas e pragmáticas.

A técnica linguística utiliza recursos que consideram características referentes à forma das palavras, como padrões de aparição de sequências de determinadas classes lexicais no corpus de estudo. Drouin (2003, p. 99) ressalta alguns padrões liguísticos observados no processo de detecção de termos:

> Essencialmente, o pressuposto é que as unidades que são particularmente frequentes, que correspondem aos padrões padrão de formação de termos (por exemplo, substantivo + adjetivo, substantivo + substantivo), ou ambos, provavelmente são termos. [14]

Assim, esses padrões de formação são levados em consideração para a detecção de termos pela técnica linguística.

14. No original: *Essentially, the assumption is that units that are particularly frequent, that correspond to standard patterns of term formation (e.g. noun + adjective, noun + noun), or both, are likely to be terms.*

Para Estopà Bagot (1999), a grande quantidade de ruído gerada (entre 55% e 75%) é um dos problemas principais dos sistemas que trabalham apenas com dados morfológicos, morfossintáticos, sintáticos e/ou lexicais. Por esta razão, afirma que pesquisadores compartilham da ideia de que o emprego de algum tipo de conhecimento semântico é a única forma de reconhecer e delimitar as unidades terminológicas de um texto especializado.

Já a técnica estatística se utiliza de informações como frequência de aparição desses itens nos corpora de estudo e de referência e a híbrida combina as duas técnicas.

Segundo Baker (2004), a análise inicial dos vocábulos de corpus é quantitativa, ou seja, probabilística/estatística, uma vez que a linguista considera que a alta chavicidade de um termo não ocorre aleatoriamente, mas é uma característica significativa para indicar a existência de um termo. Assim, por meio da comparação estatística entre o léxico de um corpus de estudo e um ou mais corpora de referência, é possível que a ferramenta analise o quanto determinada palavra é representativa em frequência relativa por meio do cálculo do que chamamos de sua chavicidade (do inglês *keyness*).

Os métodos estatísticos dependem do tamanho do corpus que utilizam, e, quanto maior o corpus, menor o silêncio, ou seja, menor a quantidade de termos não detectados devido à baixa frequência de termos no corpus. Além disso, os métodos estatísticos também geram bastante ruído, ou grande número de candidatos a termo sem valor terminológico (ESTOPÀ BAGOT, 2001).

Os sistemas baseados em conhecimento híbrido (Frantzi and Ananiadou, 1997; Dias et al., 2000) utilizam em conjunto o conhecimento estatístico e o linguístico. Tal aplicação permite aumentar a eficiência desse sistema, visto que ele condiciona os resultados apresentados (TELINE, ALMEIDA e ALUÍSIO, 2003). Esses métodos híbridos podem ocorrer de suas formas: ao aplicar primeiro o conhecimento estatístico e depois o linguístico, e ao utilizar a estatística apenas como um complemento do conhecimento linguístico. Utilizado da primeira forma, acontecem os mesmos problemas em relação ao silêncio encontrados nos sistemas apenas estatísticos. Quando o método híbrido é aplicado da segunda forma, observa-se uma melhora nos resultados causada pelo auxílio da estatística na detecção de um termo.

Porém, por mais que o método utilizado gere uma lista de termos "prováveis" organizados por relevância, resta decidir quais são suficientemente representativos de determinado domínio. Por isso, Baker (Ib. idib.) indica que a próxima etapa da análise desse corpus deve ser qualitativa, ocorrendo a partir da análise humana dos resultados, como das causas da chavicidade e do sentido dos candidatos a termos fornecidos pela ferramenta dentro do contexto. Sobre essa relação entre análise quantitativa e qualitativa, Baker afirma que dados estatísticos servem apenas para atrair a atenção do analista para aspectos que provavelmente devem ser analisados mais de perto.

Apesar de sua praticidade, segundo Conrado et al. (2020), e do fato de existir por mais de 20 anos, a EAT ainda apresenta alguns problemas a serem solucionados, sendo os maiores: 1. termos extraídos que não são verdadeiramente termos (ruído) ou termos que deixam de ser extraídos (silêncio); 2. dificuldade de se processar um número grande de candidatos a termos; 3. processo de validação manual de termos, que requer tempo e mão de obra especializada; 4. falta de consenso entre especialistas sobre o que é e o que não é um termo.

Mas qual seria o melhor método para extração de termos: o manual ou o automático? No artigo de Teline, Almeida e Aluísio (2003), as autoras comparam os processos manual e automático de extração de termos por meio do projeto intitulado ExPorTer (http://www.nilc.icmc.usp.br/nilc/projects/termextract.htm). No processo manual de extração de termos, foi privilegiado o critério semântico; no processo automático, o critério de frequência.

Em relação à extração manual dos candidatos a termos do corpus, as pesquisadoras ressaltam as dificuldades encontradas para se distinguir uma palavra (unidade da língua geral) e um termo (unidade das comunicações especializadas). Elas enfatizam também a necessidade do acompanhamento de especialistas no processo, a quem

> cabe sugerir as fontes relevantes e mais representativas para servir de base para a constituição do corpus, como também apontar, nas listas de candidatos a termo elaboradas pelo terminólogo, os termos que devem ser incluídos e os que devem ser rechaçados. (TELINE, ALMEIDA e ALUÍSIO, 2003, p. 2)

Entretanto, as autoras apontam que essas dificuldades podem ser minimizadas com o auxílio de uma ferramenta que possibilite a extração de termos de forma automática, ao considerar o critério frequência para a seleção dos candidatos a

termo. Assim, sugerem que, após realizada a extração automática, as listas de candidatos a termo devem ser enviadas aos especialistas da área após inseridos nos campos e/ou subcampos da árvore de domínio de forma a facilitar a validação. Assim, considera-se o critério semântico como complemento ao critério de frequência.

2.3.1
APLICAÇÃO DE PONTOS DE CORTE

Uma vez que quanto menor a chavicidade, menor a probabilidade de os itens lexicais extraídos pelas ferramentas serem realmente termos, a aplicação de pontos de corte, tomando padrões estatísticos para eliminar candidatos não tão representativos, pode ajudar nessa análise. No estudo realizado por Lopes e Vieira (2013), foi realizada uma comparação entre pontos de corte para listas de termos extraídos. Os seguintes tipos de pontos de corte foram analisados: absolutos, por limiar e relativos.

O ponto de corte absoluto, considerado a técnica mais simples, consiste em selecionar um número arbitrário de termos. Sugere-se, ainda, que as listas sejam analisadas separadamente de acordo com o número de elementos dos termos (unigramas, bigramas, trigramas etc.), escolhendo um ponto de corte para cada uma dessas listas, conforme realizado em diversos estudos (YANG e CALLAN 2008, LOPES et al. 2009b, EVERT 2010, DING et al. 2011, apud LOPES e VIEIRA, 2013).

Uma outra forma muito utilizada de descarte de termos e de uso de pontos de corte se dá por meio da determinação de limiares arbitrários entre as ocorrências encontradas no corpus. Bourigault e Lame (2002, apud Lopes e Vieira, 2013), por exemplo, sugerem descartar os termos que tenham menos de 10 ocorrências. Esse índice escolhido representa a frequência absoluta de termos. Dessa forma, considera termos relevantes utilizando um ponto de corte baseando-se em determinado limiar, por meio do qual considera apenas os termos cujo índice de ocorrência fica acima do limiar escolhido.

A última alternativa para pontos de corte analisada, conforme proposta por Maynard et al.(2008 apud LOPES e VIEIRA, 2013) foi a de ponto de corte relativo. Aqui, sugere-se reter apenas um percentual da lista de termos extraídos, conforme considerado proporcional, da seguinte forma: primeiro, aplica-se um índice em que

calcula a relevância de um termo no corpus de estudo em contraste com sua relevância em um corpus contrastante – o índice *tf-dcf* (LOPES et al., 2012). Em seguida, considera-se apenas 15% da lista extraída (valor em que foi obtido o melhor limiar para termos simples e complexos) e excluem-se os termos onde o índice *tf-dcf* é inferior ao limiar 2.

Cabe relembrar, porém, que, mesmo após aplicado o ponto de corte, a lista gerada deve passar posteriormente por validação humana, a fim de verificar se são realmente termos dentro da área de especialidade analisada.

2.3.2
FERRAMENTAS E ANÁLISE DE CORPORA

Existem inúmeras ferramentas de análise de corpora. Com o crescimento da quantidade de textos em formato eletrônico e a melhoria da velocidade e da conectividade da Web, as ferramentas que precisam ser instaladas e gerenciadas direto do computador do usuário foram perdendo a popularidade entre lexicógrafos que lidam com grandes projetos, já que o corpus e a manutenção do software requerem memória e computadores robustos. Consequentemente, a maioria dos projetos lexicográficos hoje em dia prefere ferramentas de corpus online que utilizam protocolos *http*, de forma que os usuários não precisem mais instalar softwares nos seus computadores e podem operar com muitos mais dados, como até bilhões de palavras (KILGARRIFF e KOSEM, 2013).

Teixeira (2010) realizou um estudo comparando quatro desses aplicativos de processamento e análise de corpus online – Corpógrafo 4.0, e-Termos, WordSmith Tools 3.0 e ZExtractor – a fim de avaliar qual deles apresentava melhor índice de acerto na extração de termos. Para o estudo, a autora utilizou um corpus de cerca de 500.000 palavras e avaliou a acuidade na extração de unigramas (termos simples). Dentre essas ferramentas, o Corpógrafo foi o que apresentou o melhor resultado, liderando o ranking com 27.56% de acerto, seguido, respectivamente, pelo ZExtractor (com 26.05%), Wordsmith Tools 3.0 (com 21,77% de acerto) e, por último, pelo e-Termos (com 14,44%).

O Corpógrafo[16], atualmente na versão 5, é uma ferramenta disponível na Web de gestão e pesquisa de corpora, resultado de um trabalho multidisciplinar de

16. https://www.linguateca.pt/Corpografo/

pesquisadores da Linguateca, da Faculdade de Letras da Universidade do Porto (FLUP), Portugal.

Outra ferramenta que vale ser mencionada é a OntoLP, cujo uso para a extração de termos de um corpus da área de Pediatria foi descrito no artigo de LOPES et al (2007). Nesse trabalho, os resultados obtidos pela ferramenta foram comparados com os resultados de uma lista de termos de referência obtida manualmente. Esse estudo possibilitou demonstrar as vantagens do processamento com a utilização de análise sintática e semântica.

Já o pesquisador Drouin (2003) apresentou um estudo a partir da utilização de uma técnica híbrida que compara frequências lexicais em um corpus especializado e em um corpus não especializado para extrair candidatos a termos com o mínimo de ruído possível. No estudo, foi utilizado um corpus em inglês do domínio de telecomunicações dividido em três subcorpora de tamanhos diferentes (11.947, 28.583 e 8.676 palavras cada), a fim de testar os algoritmos. Por meio dessa metodologia, foi alcançada a precisão total de 81% de acerto na extração de termos, sendo 86% para termos simples e 65% para termos complexos.Esse estudo gerou a ferramenta TermoStat, então na sua primeira versão.

Observamos, assim, que tanto o Corpógrafo quanto o TermoStat se destacaram no seu desempenho. Porém, uma vez que os dois estudos (Teixeira, 2010, e Drouin, 2003) utilizaram metodologias diferentes, é difícil comparar os resultados sem testar as duas ferramentas utilizando os mesmos parâmetros. Também, apesar do melhor desempenho do Corpógrafo em relação ao das outras ferramentas testadas, foi avaliada apenas a extração de termos simples e, conforme mencionado anteriormente, os termos complexos são maioria entre os termos em um corpus especializado; já o TermoStat apresentou desempenho satisfatório para os dois tipos de termos (mais de 50% de acertos). Ainda, devido ao Corpógrafo apresentar funcionamento instável diversas vezes durante o período em que esta pesquisa foi realizada, optou-se por não utilizá-lo neste estudo.

Outra ferramenta que vem se destacando na EAT é o AntConc (ANTHONY, 2019). O programa concordanciador é disponibilizado gratuitamente para download no site de Laurence Anthony, seu criador, na internet[17]. Por meio de

17. https://www.laurenceanthony.net/software/antconc

uma interface intuitiva e amigável e uma extensa gama de funções de fácil utilização, vem ganhando popularidade não apenas para realização de pesquisas linguísticas, mas também para tradução e estudo de idiomas. O desempenho da ferramenta já foi testado com sucesso em outro estudo nosso, na versão 3.5.7 (MULLER, 2019).

Assim, optou-se, neste estudo, por testar o desempenho das duas ferramentas: o TermoStat Web 3.0 e o AntConc 3.5.7. Para isso, primeiramente, foi realizado um estudo piloto a partir de um pequeno corpus paralelo. Em seguida, foi feita a EAT propriamente dita utilizando um corpus de estudo comparável para a extração de candidatos a termo que comporiam o léxico do CFN. A seguir, descreveremos detalhadamente as duas ferramentas utilizadas.

2.3.2.1
FERRAMENTA TERMOSTAT WEB 3.0

O TermoStat Web 3.0 é um sistema híbrido que se utiliza de pistas linguísticas e de técnicas estatísticas para extrair candidatos a termos de um texto. Desenvolvido por Patrick Drouin, professor do *Observatoire de Linguistique Sens-Text* (OLST) da Universidade de Montreal, no Canadá, a ferramenta está disponível gratuitamente na internet, tem utilização online e possui layout intuitivo e amigável.

Os corpora de referência com que o programa trabalha são extensos e estão disponíveis nos idiomas espanhol, francês, inglês, italiano e português (de Portugal). O maior corpus de referência com que trabalha é o francês, com aproximadamente 28.500.000 ocorrências. O menor deles, em inglês, possui oito milhões de ocorrências, o que corresponde, em média, a 465.000 formas lexicais diferentes provenientes de artigos do jornal The Gazette, de Montreal, e do British National Corpus (BNC), publicados entre 1989 e 1989. Já o corpus de referência em português possui cerca de 10.000.000 de *tokens*, que correspondem a cerca de 542.000 formas diferentes. (TERMOSTAT, 2020).

O sistema trabalha com a análise de um arquivo único de corpus fornecido pelo usuário, que é comparado a um corpus de referência, previamente anotado. A partir dessa análise, é gerada uma lista de candidatos a termos simples e complexos, organizados por uma pontuação atribuída aos candidatos a termos a partir da reali-

zação de testes por frequência, chavicidade, *Log-likelihood* (comparação entre corpora), *Log-odds ratio* (valor proveniente de um cálculo atribuído ao candidato a termo; quanto maior o seu valor, o mais interessante ele será a partir de uma perspectiva terminológica) e *X2* (comparação das frequências de ocorrências entre candidatos a termos). O programa permite também selecionar a classe gramatical para os candidatos a termos que se deseja extrair entre adjetivos, advérbios, substantivos e verbos.

O sistema opera em três etapas: primeiramente, identifica e classifica as palavras do corpus de estudo com um anotador de classes de palavras chamado *TreeTagger*. Em seguida, o texto passa por um filtro em busca de matrizes sintáticas pré-definidas utilizando como base estruturas em que comumente se apresentam os termos.

Conforme consta em seu manual de utilização (DROUIN, 2010), as estruturas mais comuns em que os termos se apresentam são Nome; Nome + Adjetivo; Nome + Preposição + Nome; Nome + Preposição + Nome + Adjetivo; Nome + Particípio; Nome + Adjetivo + Preposição + Substantivo; Adjetivo; Advérbio e Verbo. Assim, o programa varre o corpus de estudo em busca dessas estruturas, utilizando-as como pistas para localizar os candidatos a termos em potencial.

Dessa forma, os candidatos a termos são avaliados de acordo com o teste escolhido pelo usuário para visualização de resultados, recebendo as maiores pontuações os considerados mais relevantes dentro do texto. Já os que ultrapassam o limite de aceitabilidade do sistema (*acceptability threshold*) são excluídos.

A lista dos termos selecionados após a busca oferece também os seguintes dados: *Group candidate*, com a forma que sofreu o mínimo de variações em comparação à forma lematizada fornecida pelo TreeTagger; *Frequency*, para a frequência das ocorrências; *Weight*, para a pontuação recebida pelo candidato a termo de acordo com o teste selecionado; *Spelling variants*, com as diferentes formas do candidato a termo encontradas no texto e *Matrix*, com a sequência de classes gramaticais encontradas (por exemplo, Nome + Preposição + Nome).

Além disso, cabe aqui mencionar as opções *Cloud*, com a lista de 100 termos de maior pontuação organizada em ordem alfabética; *Statistics*, que fornece o número de candidatos selecionados e o número de candidatos para cada matriz gramatical; e *Bigrams*, que apresenta as combinações de verbo + substantivo mais relevantes encontradas. Dispõe ainda de outras funcionalidades, porém este trabalho

não pretende esgotar a descrição do programa, uma vez que não utilizamos todas as suas ferramentas para a criação do léxico proposto.

2.3.2.2
FERRAMENTA ANTCONC3.5.7

O AntConc é uma ferramenta gratuita que permite realizar a análise de corpora eletrônicos. Suas inúmeras funcionalidades permitem ao pesquisador realizar uma análise aprofundada de vários documentos ao mesmo tempo, como localizar palavras-chaves (apenas unigramas) em um texto de acordo com uma grande variedade de critérios, trazendo informações sobre as palavras como frequência (organizada por um ranking) e chavicidade. Neste capítulo, não se pretende descrever todas essas funcionalidades que o programa apresenta, mas apenas aquelas que serão usadas para a criação do léxico a que este estudo se propõe.

Quanto ao tipo de documentos processados pelo programa, apesar do AntConc permitir a leitura de vários formatos como .txt, .xml e .html., sua forma mais simples de utilização é no formato txt., com a codificação UTF-8, definido pelo padrão Unicode, que abrange caracteres constantes na maioria das línguas ocidentais.

As opções avançadas do AntConc permitem ao usuário realizar uma pesquisa personalizada de acordo com a sua necessidade, com a possibilidade de inserção de uma *stoplist*, ou *palavras de parada*, que são palavras que se deseja excluir da busca, e o carregamento de qualquer corpus de referência a partir do computador do usuário, por exemplo. Permite também personalizar a busca quanto aos valores de chavicidade, oferecendo diversas opções de medidas de associação, desde estatísticas (*Chi-squared*, *Log-likelihood*), de linhas de corte etc.

A ferramenta extrai do corpus, após ajustadas as configurações pelo usuário, uma lista de palavras-chave, que nada mais são do que unigramas. O programa permite também, ao clicar no item desejado dentro da lista de palavras-chave ou ao digitá-lo na caixa de busca, localizar cada ocorrência desse item lexical dentro do contexto, possibilitando assim a análise das linhas de concordância. É possível destacar com cores a quantidade de palavras à esquerda e/ou à direita da palavra-chave dentro do contexto, conforme o ajuste do usuário. Além disso, ainda que o corpus

analisado possa ser composto por mais de um documento, o AntConc informa exatamente em qual documento cada ocorrência do candidato a termo se encontra.

Os candidatos a termos extraídos pelo programa podem integrar termos compostos e complexos. Para detectá-los, além da análise das linhas de concordância, é possível utilizar a opção *Clusters/N-Grams*, que buscará os n-grams associados às palavras-chaves detectadas pela funcionalidade *keyword*. Mais uma vez, é possível customizar a busca ao escolher a quantidade de palavras que esses n-gramas possuirão na sua composição e se os sintagmas crescerão para a esquerda ou para a direita da palavra-chave, por exemplo.

Ainda que o AntConc possua uma extensa gama de opções de ajuste, ele é de fácil utilização e permite ao usuário adaptar sua pesquisa ao seu objetivo, conseguindo assim encontrar resultados muito mais direcionados do que permitiria uma ferramenta de EAT totalmente automática como o TermoStat, por exemplo. O ponto negativo observado, conforme relatado em Müller (2019), foi que, quando utilizados *corpora* muito grandes, como os compostos por milhões de palavras, o programa trava e precisa ser reiniciado diversas vezes. Além disso, constatou-se nesse mesmo estudo que, quando os documentos são salvos em formato txt para utilização na ferramenta, algumas características da formatação original são alteradas, conforme explicitado a seguir.

> Quando convertidos para o formato txt, os textos sofrem alterações na formatação original. (..) De fato, tal prejuízo foi detectado na investigação do termo "comando e controle", por exemplo, que possui como variante no português a forma "C²", que no formato txt se transformava em C2. Não foi encontrada na versão atual do programa uma maneira de corrigir a perda ocorrida. (MÜLLER, 2019, p. 28)

Por isso, foi recomendado em Müller (Ib. Idib.) sempre recorrer ao material original quando se notar alterações nos textos ao utilizá-los no AntConc.

2.3.3
Validação de candidatos a termos extraídos automaticamente de um corpus segundo a Semântica Lexical de L'Homme (2020)

Apesar de reconhecer a importância das ferramentas computacionais que auxiliam no trabalho terminológico de identificação e extração de termos, L'Homme (2020) propõe ir um pouco mais longe e recorrer a métodos que utilizam o

ponto de vista qualitativo, ou seja, nos quais os contextos ajudam a validar ou invalidar possíveis intuições que o analisador possa ter a respeito dos itens lexicais extraídos por meio dessas ferramentas e identificados como termos em potencial.

Considerando que os termos são unidades lexicais, L'Homme (Ib. Idib.) propõe quatro critérios para a sua identificação dentro da abordagem da Semântica Lexical:

(A) a relação com um campo de conhecimento;

(B) a natureza dos argumentos;

(C) as relações morfológicas e semânticas que estabelece;

(D) as relações paradigmáticas.

Quanto ao critério A, de relação com o campo de conhecimento, L'Homme (Id. Ibid.) indica que é preciso conhecer a área de especialidade em que o termo está inserido. Este critério necessita, portanto, de conhecimento extralinguístico. É também mais facilmente aplicado a entidades como substantivos e mais dificilmente aplicado a verbos, que tendem a ser menos óbvios quanto à associação com determinada área de especialidade.

Já em relação ao critério B, identifica-se que a natureza de itens lexicais que acompanham um termo (identificados pelos critérios A, C ou D) complementando seu sentido (indicando atividades, propriedades ou relações), chamadas de "unidades lexicais predicativas" possuem características terminológicas. Logo, esses itens provavelmente correspondem também a termos. Esse critério é utilizado normalmente com verbos, mas pode ser utilizado também para substantivos, adjetivos e advérbios. Por exemplo:

empregar: **alguém**(um comandante) emprega **algo**(meios) em **algo** (em uma missão)

anfíbio: **algo** é anfíbio (ataque anfíbio)

ameaça: ameaça de **algo** (de ataque cibernético) ou **alguém** (do inimigo) a **algo**(ao Comando de Operações Navais) ou **alguém** (ao Comandante-Geral).

O critério C permite observar se a unidade lexical é morfológica e semanticamente relacionada a uma unidade lexical já identificada como um termo de acordo com os critérios A, B ou D. Nesse caso, provavelmente se tratará de um termo. Podemos utilizar como exemplo a unidade lexical *operação*, já definida

como termo no campo de conhecimento militar ao aplicarem-se os outros critérios mencionados. Nesse caso, pelo critério C, palavras derivadas morfológica e semanticamente como *operativo* e *operacional* devem ser consideradas termos.

Por último, o critério D trata da análise de candidatos a termos que estabelecem relações paradigmáticas entre uma unidade lexical já reconhecida como termo de acordo com os critérios A, B ou C. Nesse caso, a unidade lexical seria muito provavelmente um termo. Este critério abrangeria as relações paradigmáticas não abordadas pelo critério C, como *sinonímia, antonímia, meronímia* e, em certa medida, *hiperonímia*.

Segundo Cançado (2013), a relação de sinonímia existe entre palavras diferentes de significado similar (como em *face* e *rosto*, *dinheiro* e *grana*); já a relação de hiperoníma e hiponímia estaria relacionada a superordenação ou subordinação (sendo *meios* hiperônimo de *tropa* e *viatura*; e *navio* e *aeronave* hipônimo de *esquadra*). Meronímia estabeleceria uma relação de parte do todo ou contiguidade, por exemplo: *retrovisor* estabeleceria uma relação de meronímia com *viatura*.

2.4
MÉTRICAS

Neste estudo, utilizamos como índices as medidas de precisão, abrangência e medida F (van RIJSBERGEN, 1975, apud LOPES e VIEIRA, 2013), consideradas tradicionais na área de recuperação de informação, para comparação das listas de termos extraídas pelas ferramentas testadas. Ainda, assim como Lopes e Vieira (Ib. Idib.), utilizaremos a denominação LR (lista de referência), considerando os termos utilizados no estudo piloto, e LE (lista extraída) para a lista de termos extraídos pela ferramenta.

A precisão (*P*) é calculada pela razão entre o número de termos da lista de referência que constam na lista extraída pela ferramenta, representando o tamanho da intersecção entre a LR e a LE, e o tamanho da lista de todos os termos que foram extraídos e considerados (| LE |). Logo, segundo Lopes e Vieira, "a precisão expressa o percentual de termos corretamente extraídos, ou seja, o percentual dos termos localizados como corretos, quantos são efetivamente corretos" (2013, p. 82).

Já a abrangência (*R*) representa a razão entre o número de positivos da LE que constam na lista de referência (LR) e o número total de termos da lista de referência (| *LR* |). Logo, "a abrangência expressa o percentual de termos da lista de referência coberta pela extração de termos feita" (Ib. Idib., p. 82).

A medida F (*F*) é uma expressão numérica que representa o equilíbrio entre *P* e *R*; ou seja, é a média harmônica entre os valores de precisão e abrangência. Lopes e Vieira apontam que "os valores da medida F são valores situados entre P e R e quanto maior for a diferença entre esses valores, mais próxima a medida F será do menor valor entre eles. (2013, p. 82). Essa medida também pode ser utilizada como ponto de corte em um corpus.

Portanto, as seguintes equações mostradas na figura a seguir serão usadas para a realização dos cálculos:

$$P = \frac{|LR \cap LE|}{LE} \quad R = \frac{|LR \cap LE|}{LR} \quad F = \frac{2 X P X R}{P+R}$$

Figura 6 – Equações para cálculo de precisão (P), abrangência (R) e medida F (F)

Para Lopes e Vieira,

> O uso desses índices de qualidade é bastante difundido em diversas áreas, e.g. [Boreczky and Rowe 1996, Thomas et al. 2000, Fernandes et al. 2010]. Na área de PLN, e em especial nas tarefas de extração de termos, diversos trabalhos justificam a sua validade baseados em seus resultados numéricos, e.g., [Manning and Schutze 1999, ¨ Bell et al. 1999, Hulth 2004, Lopes et al. 2010b]". (2013, p. 82)

Logo, por meio desses cálculos, e dispondo de uma lista de referência, é possível comparar o desempenho das ferramentas em análise e descobrir qual delas apresenta o melhor índice de precisão, abrangência e o de equilíbrio entre as duas medidas – a medida F, tanto para termos simples, como para termos compostos. No caso do TermoStat, também serão comparados os mesmos índices em relação à EAT nos idiomas inglês e português, uma vez que o programa apresenta o próprio corpus de referência.

3
MATERIAL E METODOLOGIA

A metodologia deste trabalho compreende duas fases. Na primeira fase, foi feito um estudo piloto com o objetivo de criar uma lista de referência e, consequentemente, avaliar a qualidade das ferramentas de EAT. Na segunda fase, os resultados obtidos com o estudo piloto foram aplicados a um novo corpus, com o objetivo de criar o léxico *Espírito de Corpus*.

Na primeira fase, para a seleção e utilização dos corpora de estudo, foi seguida a metodologia proposta por Tognini-Bonelli (2001). Antes de iniciada a EAT propriamente dita, foi feito um estudo piloto a partir da análise automática e manual de um corpus paralelo a fim de construir uma lista de referência e avaliar o desempenho das ferramentas TermoStat Web 3.0 e AntConc 3.5.7 tanto quanto à extração de termos simples e compostos/complexos quanto à EAT nos idiomas inglês e português. Esse teste foi realizado conforme as métricas de avaliação de desempenho precisão, abrangência e medida F, apresentadas no capítulo 2.

O quadro a seguir traz as principais características das ferramentas avaliadas, conforme descritas com mais detalhes na seção 2.3.3.

AntConc 3.5.7	**TermoStat Web 3.0**
• Disponível gratuitamente;	• Disponível gratuitamente;
• Sistema estatístico;	• Sistema híbrido (linguístico e estatístico);
• Precisa fazer download;	• Ferramenta online;
• Permite carregar corpus com mais de um arquivo;	• Permite baixar corpus de um único arquivo;
• Permite aplicar *stoplist*;	• Não permite aplicar *stoplist;*
• Permite utilizar uma série de medidas de associação;	• Permite utilizar uma série de medidas de associação;
• Permite inserir corpus de referência;	• Corpora de referência próprios e anotados em cinco idiomas;
• Detecta *nodes* (unigramas) e, a partir deles, seus colocados a partir da análise das linhas de concordância.	• Detecta unigramas e n-gramas.

Quadro 3 - Características do AntConc 3.5.7 x TermoStat Web 3.0

Conforme podemos observar no quadro 5, ambas as ferramentas são disponíveis gratuitamente e permitem realizar a análise de corpora a partir da utilização

de diversas medidas de associação. Enquanto o AntConc possui um sistema estatístico de análise de corpora e possui diversas funcionalidades ajustáveis, além de possibilitar a aplicação de *stoplists* e a inserção de corpora de referência do arquivo do próprio usuário, o TermoStat utiliza um sistema linguístico e estatístico de análise de corpora e não permite o ajuste manual de suas configurações nem a aplicação de *stoplists*. Utiliza também seu próprio corpus de referência, anotado e disponível em 5 idiomas, entre eles, português e inglês.

Após o estudo piloto, foi realizada a análise híbrida (automática e manual) de um corpus maior, bilíngue e comparável para a confecção do léxico. Para a validação e análise manual dos termos extraídos automaticamente, foram aplicados os critérios de validação de termos segundo a proposta de L'Homme (2020). Já para encontrar os equivalentes terminológicos na língua meta, foi utilizada a metodologia proposta por Tognini-Bonelli (2001). A análise dos termos contou com o apoio também de material terminológico de referência na área.

Em suma, foram seguidas as seguintes etapas metodológicas:

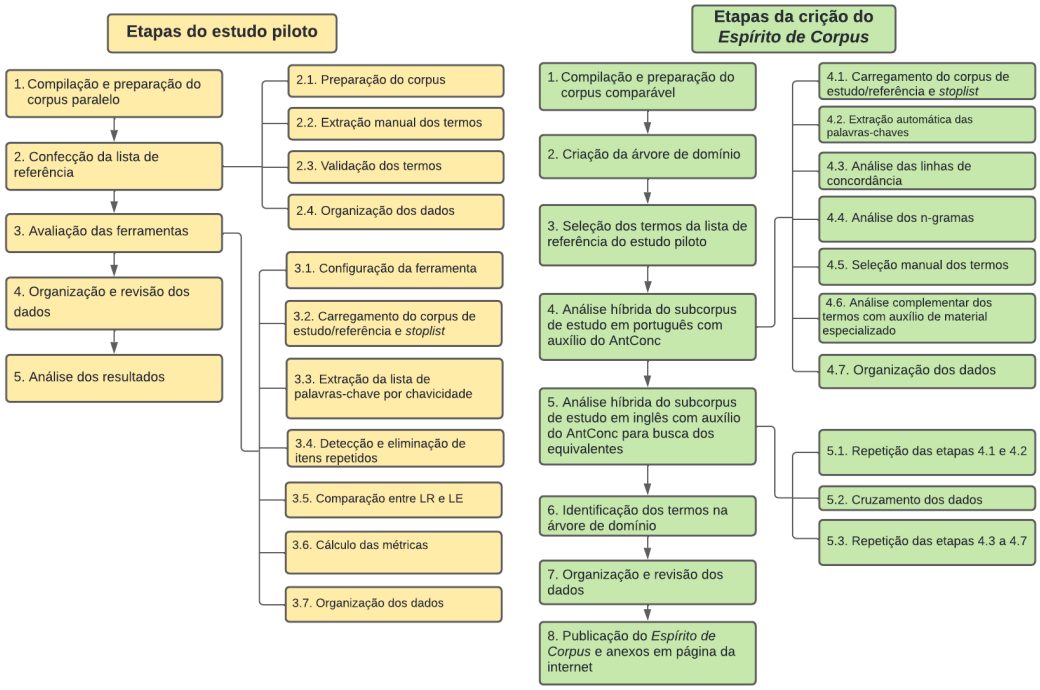


Figura 7 - Etapas metodológicas da pesquisa

Essas etapas serão explicadas detalhadamente nos próximos capítulos ou seções deste trabalho.

3.1
OS CORPORA UTILIZADOS

3.1.1
O CORPUS PARALELO

Primeiramente, a fim de avaliar as ferramentas AntConc e TermoStat no estudo piloto, foi compilado um pequeno corpus paralelo. O texto desse corpus foi escrito originalmente em português, com a sua respectiva tradução para o inglês; logo, é composto por dois subcorpus, um em cada idioma. As características desse corpus podem ser vistas no quadro a seguir:

	CORPUS PARALELO (BILÍNGUE)	
	SUBCORPUS DE ESTUDO L1 (PORTUGUÊS)	SUBCORPUS DE ESTUDO L2 (INGLÊS)
Nome	APS Base FFE_Pt	APS Base FFE_En
Ano	2021	2021
Textos	01. CORPO DE FUZILEIROS NAVAIS - Força Estratégica de Pronto Emprego e de Cartáter Anfíbio e Expedicionário	01. *BRAZILIAN MARINE CORPS - National Strategic Amphibious Expeditionary Force in Readiness*
Tipologia	Apresentação de slides	Apresentação de slides
Assunto	Apresentação sobre o Corpo de Fuzileiros Navais (história, estrutura, tarefas e emprego)	Apresentação sobre o Corpo de Fuzileiros Navais (história, estrutura, tarefas e emprego)
Fonte	Arquivo do Comandante da Força de Fuzileiros da Esquadra	Arquivo do Comandante da Força de Fuzileiros da Esquadra
Número de palavras	1.356	1.261

Quadro 4 - Corpus paralelo utilizado no estudo piloto

O corpus paralelo consiste em uma apresentação de slides realizada para jornalistas de mídias especializadas em defesa militar apresentando o CFN. Sua tradução foi realizada a fim de apresentar o CFN em uma palestra ocorrida durante uma inspeção da Organização das Nações Unidas realizada no Complexo Naval da Ilha do Governador, em 19 de julho de 2021. O autor desses slides, assim como o tradutor, foi o Vice-Almirante (Fuzileiro Naval) Carlos Chagas, atual Comandante da Força de Fuzileiros da Esquadra (FFE).[18]

18. Vale ressaltar que o Vice-Almirante Carlos Chagas é Doutor em Relações Internacionais pela PUC-Rio e Mestre em Estudos Militares pela *United States Marine Corps University Quantico*, em Virginia, EUA (BRASIL, 2021d).

O documento recebido, inicialmente, consistia em um arquivo em pdf com 134 slides intercalados entre sua versão em português e em inglês. Primeiramente, o texto da apresentação de slides foi extraído na íntegra e transformado no formato .txt. A partir daí, foram criados dois arquivos separados por idiomas: Aps CFN_Pt.txt e Aps CFN_En.txt. Esse formato foi escolhido por ser compatível para leitura por ambas as ferramentas em estudo. Observa-se no quadro 3 que o texto em português é um pouco mais extenso do que o texto em inglês, pois possui maior variedade de itens lexicais quando comparados a sua tradução em inglês. Por exemplo: Batalhão de Artilharia / BTL de Artilharia = Marine Artillery Battalion. Componente de Apoio de Serviço ao Combate / CASC = Logistics Combat Element.

3.1.2
O CORPUS COMPARÁVEL

Em seguida, para a confecção do léxico, foi compilado um corpus comparável bilíngue, cerca de 67 vezes maior do que o corpus paralelo, que contém as especificações descritas no quadro a seguir.

	CORPUS COMPARÁVEL (BILÍNGUE)	
	SUBCORPUS DE ESTUDO L1 (PORTUGUÊS)	*SUBCORPUS* DE ESTUDO L2 (INGLÊS)
Nome	Manuais do CFN	Manual do USMC
Ano	2020	2020
Textos	1. CGCFN-1003 MANUAL BÁSICO DO FUZILEIRO NAVAL 2. CGCFN-3101.1 MANUAL DO PELOTÃO DE FUZILEIROS NAVAIS	1. *MCIP 3-10A.4i* *Marine Rifle Squad*
Tipologia	Manual	Manual
Assunto	História, tradições e explanações sobre as Operações Anfíbias, assim como organização, tarefas e emprego tático do Pelotão de Fuzileiros Navais no combate	Organização, tarefas e emprego tático do Pelotão de Fuzileiros Navais no combate
Fonte	Comando-Geral do Corpo de Fuzileiros Navais https://www.marinha.mil.br/cgcfn/	United States Marine Corps University Research Library https://grc-usmcu.libguides.com/library-of-the-marine-corps/
Número de palavras	87.568	80.649

Quadro 5 - Corpus comparável utilizado para a confecção do léxico

Conforme se pode inferir ao comparar as características dos subcorpora acima descritos, os documentos que compõem o corpus possuem ano de publicação (2020) e tipologia (manual) idênticas, assim como tamanho (de 80 a 87 mil pala-

vras) e conteúdo bastante similares. Quanto ao assunto tratado nesse corpus, ressalta-se que os dois subcorpora abordam a organização, tarefas e emprego tático do Pelotão de Fuzileiros Navais em combate, mas apenas a versão em português traz informações a respeito da história e tradição da força brasileira. Observando a figura 8, pode-se perceber a similaridade do conteúdo entre os manuais em português, do CFN, e em inglês, do USMC.

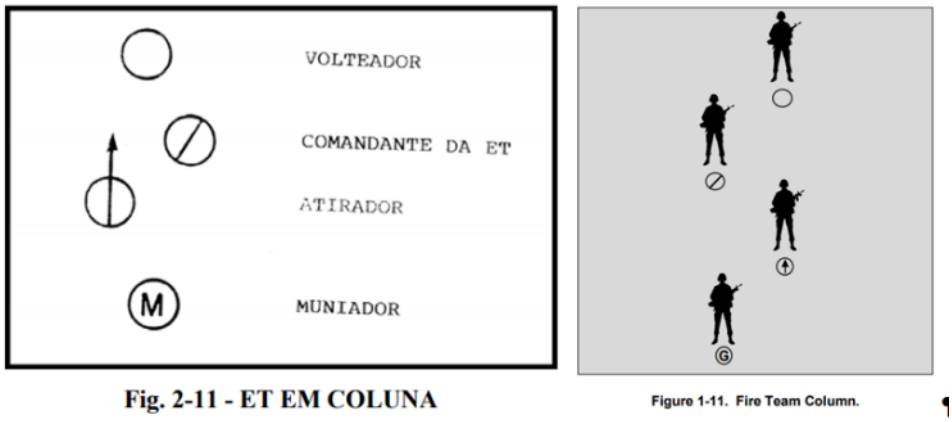


Figura 8 – Comparação de imagens extraídas dos dois subcorpora. Fonte: Brasil, 2020a, item 2-8; e u.s. Marine Corps, 2020, p. 24, respectivamente.

Dessa forma, ainda que não tenha sido possível compilar um corpus paralelo, conforme propõe Tognini-Bonelli (2001) como melhor hipótese para se extrair terminologia bilíngue, a similaridade de conteúdo dos materiais permite localizar facilmente os equivalentes terminológicos, dependendo menos da intuição e do conhecimento prévio do analisador.

3.1.3
O CORPUS DE REFERÊNCIA

Como corpora de referência, foram utilizados extratos dos corpora CETEN Folha (2018) para o português e do TIME Magazine Corpus of America (s.d.) para o inglês, ambos com cerca de 1.100.000 palavras e compostos por textos jornalísticos. Cada corpus de referência é cerca de 13 vezes maior do que o subcorpus de estudo de cada idioma; logo, permite estabelecer um contraste com o corpus especializado a fim de identificar, por meio de uma análise quantitativa, candidatos a termos de alta chavicidade, conforme proposto por Baker (2004).

4
ESTUDO PILOTO

Este capítulo visa a descrever um estudo realizado com o objetivo de avaliar duas ferramentas de extração de candidatos a termos: o TermoStat Web 3.0 e o AntConc 3.5.7.

Até onde sabemos, não há hoje, para a língua portuguesa, um corpus e seu respectivo léxico padrão ouro disponíveis que permitam avaliar ferramentas de extração automática de termos (EAT) quanto a precisão e abrangência.

Apesar de já haver estudos que avaliam a precisão de ferramentas de EAT, como o trabalho de Teixeira (2010), é difícil avaliar o desempenho de uma ferramenta quanto à abrangência devido à necessidade de um material especializado. Essa lista de termos que serve como gabarito normalmente é compilada por terminólogos, que podem levar em conta não só dados de um corpus, mas também de outras fontes. Há também glossários feitos a partir de um corpus analisado com o auxílio de ferramentas, os quais normalmente se limitam a avaliar as listas de termos fornecidas por ferramentas de extração de candidatos a termos. Nesse caso, dificilmente os pesquisadores examinam detalhadamente o corpus em busca de termos que não tenham sido identificados, ou seja, dos falsos negativos que tenham sido deixados de lado pela extração automática (e que são capturados pela abrangência), e o trabalho com as ferramentas consiste justamente em eliminar a leitura convencional de um corpus em busca de candidatos a termos.

Assim, esse gabarito ou léxico padrão ouro, que aqui chamaremos de lista de referência (LR), é útil tanto para o cálculo da precisão quanto da abrangência da ferramenta, apesar de não ser imprescindível no primeiro caso.

Dessa forma, a intenção deste estudo piloto, feito com um corpus pequeno, que pode ser analisado manualmente na íntegra, da área temática do CFN, é comparar os resultados da EAT das duas ferramentas em análise com a LR confeccionada a partir do mesmo corpus. Assim, pretendemos verificar qual das duas ferramentas possui os melhores índices de precisão, abrangência e medida F, considerando os seguintes fatores: idioma (português/inglês) e tipos de termos (unigramas e compostos/complexos). Também será analisada a distribuição dos termos nas listas fornecidas pelas ferramentas a fim de verificar a utilidade da aplicação do corte.

Ao final do estudo piloto, a ferramenta com o melhor desempenho será selecionada para a extração de termos do léxico do Corpo de Fuzileiros Navais.

As duas ferramentas foram avaliadas com aplicação de linha de corte o mais aproximada possível e, no caso do AntConc, também foi feita a avaliação sem linha de corte, uma vez que apenas essa ferramenta das duas utilizadas permite essa configuração. Para a detecção dos termos simples (unigramas) em cada uma das ferramentas, optamos por extrair todas as palavras-chave formadas por um único item lexical, independente da classe gramatical. Para os termos compostos e complexos, foram considerados os bigramas, trigramas e quadrigramas extraídos pelas ferramentas como palavras-chave, assim como foi feito na LR. Uma vez que o TermoStat se propõe a entregar termos simples e complexos em sua lista de palavras-chave, foram analisadas todas as palavras-chave fornecidas por ela e, caso o termo estivesse incompleto, aquela entrada era desconsiderada, ou seja, era considerada um erro da ferramenta. Considerando que o AntConc, após o teste de chavicidade, fornece em sua lista de palavras-chave apenas unigramas, conforme já descrito no capítulo 2.3.2.2., para detectar os termos compostos/complexos foi utilizada a opção *Clusters/N-Grams*, que buscará os n-gramas associados às palavras-chaves detectadas pela ferramenta keyword.

O estudo piloto seguiu as seguintes etapas explicitadas na figura 9[19], que serão explicadas com mais detalhes a seguir:

19. A figura 9 é um extrato da figura 7, apresentada no capítulo 3.

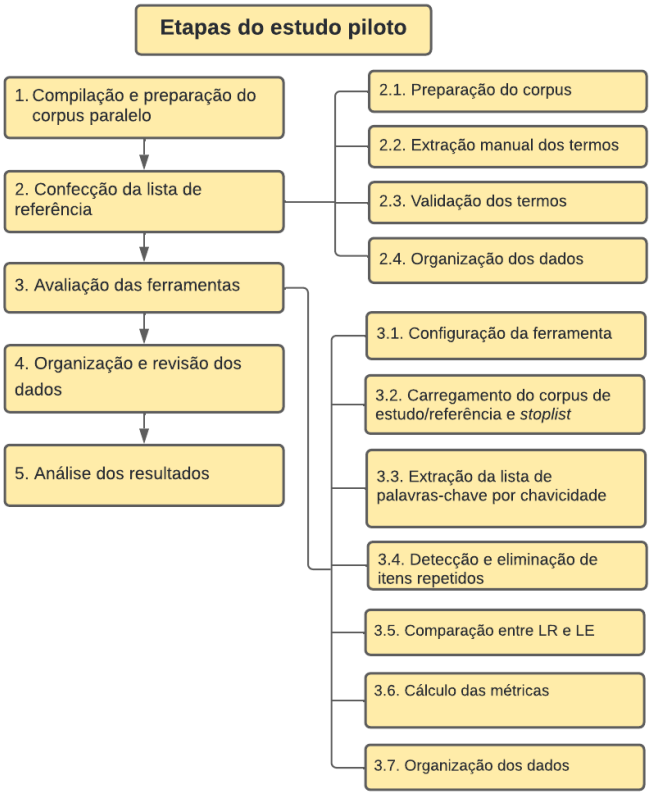


Figura 9 – Etapas do Estudo Piloto

Uma vez que as informações sobre os corpora já foram disponibilizadas na seção 3.1, serão detalhadas a seguir as etapas subsequentes à compilação do corpus paralelo.

4.1
Criação da lista de referência

A fim de criar a nossa LR, primeiramente foi realizada a análise manual minuciosa de um pequeno corpus paralelo composto por textos escritos originalmente em português, pertencentes à subárea de estudo do CFN, e sua respectiva tradução para o inglês. Desse corpus, foram extraídos manualmente os termos nas duas línguas de estudo, o que possibilitou, posteriormente, testar as duas ferramentas de EAT mencionadas ao comparar as listas de termos extraídos sem e com o auxílio da máquina.

Para a criação do léxico que originou a LR, buscou-se identificar a terminologia da área militar existente no corpus paralelo. Aqui, optou-se por não se limitar

à seleção de termos apenas da subárea do CFN, uma vez que ainda não há um material especializado (glossários, dicionários etc.) apenas desse subcampo de domínio para validação dos termos. Vale acrescentar que alguns termos que não estavam contidos no material especializado, mas que fazem parte da terminologia militar e constavam no corpus de estudo, também foram inseridos na LR, dos quais trataremos mais à frente.

Para a extração da terminologia militar do corpus de estudo, foram seguidas as etapas explicitadas a seguir.

4.1.1
EXTRAÇÃO MANUAL DOS TERMOS

Em seguida, todos os termos da área de estudo constantes no corpus foram identificados manualmente pela especialista. A lista resultante dessa identificação gerou 257 itens em português e 156 itens em inglês. Foi observado que a apresentação de slides em português era um pouco mais longa; logo, a versão em português possuía mais termos, como nomes de armamentos e de algumas organizações militares.

As duas listas geradas nos idiomas de estudo foram, em seguida, inseridas em planilhas no Excel, organizadas por relação de equivalência. É importante observar que essa relação de equivalência não importa para a avaliação das ferramentas, mas esse tipo de organização da informação facilita a posterior confecção das entradas do léxico.

4.1.2
VALIDAÇÃO DOS TERMOS

Esses termos foram, então, buscados no material terminológico especializado em terminologia militar (BRASIL, 2015; US ARMY, 2015) a fim de verificar se realmente constavam como termos pertencentes a esse campo de estudo e, assim, se podiam permanecer na LR. Porém, alguns itens lexicais que não existiam no material especializado foram, ainda assim, adicionados à lista de referência (em sua maioria, referentes a nomes de organizações militares, como Base da Ilha das Flores, Base da Ilha do Governador etc.), assim como suas abreviações, além de nomes de operações militares (Operação Verde Brasil, por exemplo), os quais, por serem

substantivos próprios, não integram normalmente glossários e dicionários, que, em sua maioria, listam apenas substantivos simples. Isso foi feito em relação a verbos (manobrar, operar) e adjetivos (naval).

Não só os termos previamente selecionados foram validados, mas alguns termos considerados pouco específicos, e, por isso, que não foram selecionados na análise manual, passaram a compor a LR após consulta ao material especializado. Alguns candidatos a termos eliminados da LR foram *área de apoio, crimes transnacionais* e *Zona Econômica Exclusiva (ZEE)*; alguns exemplos de itens acrescentados à lista são *pessoal*, *material*, *terra* e *Águas Jurisdicionais Brasileiras*.

Para melhor comparação com os resultados das ferramentas, foram mantidos na LR apenas os termos compostos por até quatro palavras (quadrigramas). Os termos maiores foram acrescentados a uma lista à parte para posteriormente serem integrados ao léxico. Alguns exemplos desses termos são *Comando-Geral do Corpo de Fuzileiros Navais* e *Comando da Força de Fuzileiros da Esquadra*.

4.1.3
ORGANIZAÇÃO DOS DADOS

Em seguida, as listas de referência foram organizadas em ordem alfabética, a fim de facilitar a visualização dos termos nas etapas posteriores, conforme a figura a seguir.

	A	B	C	D	E	F	G
1		TERMOS SIMPLES_PT		TERMOS COMPLEXOS_PT - ATÉ QUADRIGRAMA	TERMOS SIMPLES_EN		TERMOS COMPLEXOS_EN
2	1	ADSUMUS	1	1 Btl Infantaria	AAV	1	1 INFANTRY BN
3	2	AMEAÇA	2	2º Btl Infantaria	ADSUMUS	2	1ST RIVERINE OPERATIONS BATTALION
4	3	ANFÍBIO	3	3 Btl Infantaria	AMPHIBIOUS	3	2 º INFANTRY BN
5	4	ATAQUE	4	AÇÃO CÍVICO-SOCIAL	ARMY	4	2ND RIVERINE OPERATIONS BATTALION
6	5	BATALHÃO	5	ÁGUAS JURISDICIONAIS BRASILEIRAS	ATTACK	5	3RD INFANTRY BATTALION
7	6	BTL	6	AMAZÔNIA AZUL	BATTALION	6	3RD RIVERINE OPERATIONS BATTALION
8	7	CASC	7	ASSISTÊNCIA HUMANITÁRIA	BN	7	AIR COMBAT BATTALION
9	8	CCA	8	ATIVIDADE BENIGNA	BOARDING	8	AIR COMBAT ELEMENT
10	9	CCT	9	BASE DO RIO MERITI	CBRN	9	AMPHIBIOUS AND EXPEDITIONARY FORCE
11	10	CFN	10	BATALHÃO DE ARTILHARIA	COMBAT	10	AMPHIBIOUS ASSAULT VEHICLE
12	11	CIAB	11	BATALHÃO DE BLINDADOS	COMMANDANT	11	AMPHIBIOUS DIVISION
13	12	COMANDANTE	12	BATALHÃO DE COMBATE AÉREO	COMMANDER	12	AMPHIBIOUS DIVISION COMMAND
14	13	COMBATE	13	BATALHÃO DE ENGENHARIA	DEFENSE	13	AMPHIBIOUS OPERATION
15	14	DEFESA	14	BATALHÃO DE OPERAÇÕES ESPECIAIS	DOCTRINE	14	BASE RIO MERITI
16	15	DESEMBARQUE	15	BATALHÃO DE VIATURAS ANFÍBIAS	EMPLOY	15	BENIGN ACTIVITY
17	16	DOUTRINA	16	BATALHÃO LOGÍSTICO	EMPLOYMENT	16	BLUE AMAZON
18	17	EFETIVO	17	BATALHÕES DE INFANTARIA	EXPEDITIONARY	17	BOARDING TROOP
19	18	EMA-302	18	BATALHÕES DE OPERAÇÕES RIBEIRINHAS	GCE	18	BRAZILIAN JURISDICTIONAL WATERS
20	19	EMBARCAÇÃO	19	BRIGADA ANFÍBIA	LAND	19	BRAZILIAN NAVY
21	20	EMBARQUE	20	Btl Comando e Controle	LANDING	20	BRAZILIAN NAVY AMPHIBIOUS CONJUGATE
22	21	EMPREGAR	21	Btl de Artilharia	LCE	21	CBRN Defense Center
23	22	EMPREGO	22	Btl de Blindados	MAGTF	22	CHIEF OF NAVAL OPERATIONS
24	23	ENC	23	Btl de Combate Aéreo	MANEUVER	23	CIVIC-SOCIAL ACTION
25	24	EXÉRCITO	24	BTL DE DEFESA NBQR	MARINE	24	COAST GUARD
26	25	EXPEDICIONÁRIO	25	Btl de Engenharia	MARITIME	25	COMAND AND CONTROL BATTALION
27	26	FORÇA	26	Btl de Infantaria	MATERIEL	26	COMBATANT PERFORMANCE LAB
28	27	GPTOPFUZNAV	27	Btl de Logística	MILITARY	27	COMMAND AND CONTROL
29	28	GRUPAMENTO	28	Btl de Operações Especiais	MISSION	28	COMMAND ELEMENT
30	29	GUERRA	29	Btl de Operações Ribeirinhas	NAVAL	29	COMMANDER OF THE NAVY
31	30	MANOBRA	30	Btl de Viaturas Anfíbias	NEO	30	CYBER WARFARE
32	31	MANOBRAR	31	CENÁRIO ESTRATÉGICO DE INTERESSE	OPERATING	31	DISASTER RELIEF OPERATION
33	32	MATERIAL	32	CENTRO DE DEFESA NBQR	OPERATION	32	DISTRICT FORCE
34	33	MEIOS	33	CIA DE POLÍCIA	OPERATIONAL	33	FLEET MARINE FORCE
35	34	MISSÃO	34	COMANDANTE DA MARINHA	PERSONNEL	34	FLEET MARINE FORCE COMMAND
36	35	NAVAL	35	COMANDANTE DE OPERAÇÕES NAVAIS	PIRACY	35	FLEET MARINE FORCE COMMANDER

Figura 10 - Extrato das LR em português e inglês

A figura 10 é um extrato da tabela das LR, cujos termos em português e inglês estão organizadas lado a lado, em ordem alfabética. As LR na íntegra podem ser encontradas nos Apêndices 1 a 4.

Conforme é possível observar, as células que continham termos simples foram coloridas de amarelo, e as células que continham termos complexos foram coloridas de laranja. a tabela a seguir mostra o total de termos manualmente identificados em cada idioma.

TERMOS DA LISTA DE REFERÊNCIA					
Português			**Inglês**		
Unigramas	Bigramas, trigramas e quadrigramas	**TOTAL**	Unigramas	Bigramas, trigramas e quadrigramas	**TOTAL**
50	98	148	51	107	158

Tabela 1 – Quantidade de itens lexicais das LR

Prontas as LR, foi a vez de realizar os testes com as ferramentas, cujas etapas serão descritas a seguir.

4.2
Avaliação das ferramentas

Devido às peculiaridades de cada ferramenta, explicitadas no quadro 5, sendo que uma fornece uma busca muito mais automatizada, com corpus de referência próprio e que permite mínimos ajustes realizados pelo usuário (TermoStat) e a outra em que os ajustes são inseridos de forma quase que totalmente manual, o que permite ao usuário um controle muito maior dos resultados (AntConc), procurou-se ajustar a segunda com uma configuração que se aproximasse ao máximo da primeira, de modo a permitir uma comparação o mais adequada possível da eficácia das duas ferramentas. Por isso, foi escolhido realizar primeiro o teste com o TermoStat, a fim de analisar seus resultados de acordo com suas configurações automáticas para, assim, ajustar as configurações do AntConc, com o intuito de alinhar o funcionamento das duas ferramentas o máximo possível.

A seguir, explicaremos mais detalhadamente como ocorreu cada uma das etapas mencionadas para avaliar cada uma das ferramentas com os subcorpora do corpus de estudo em português e em inglês, conforme enumeradas na figura 9.

4.2.1
Avaliação do TermoStat WEB 3.0

Para os testes com a ferramenta TermoStat, primeiramente, foi feito o acesso com nome de usuário e senha no seu respectivo site. Nesse ponto, a interface da ferramenta pôde ser configurada para utilização em inglês ou francês. Após entrada, foi feito o carregamento e a seleção do subcorpus, e as seguintes especificações foram selecionadas: idioma (português ou inglês, de acordo com o idioma do corpus carregado para alinhamento do corpus de referência), termos simples e complexos e todas as categorias gramaticais, conforme especificados na figura a seguir.

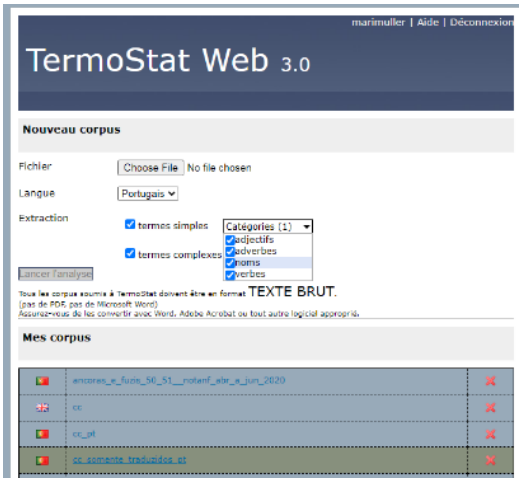
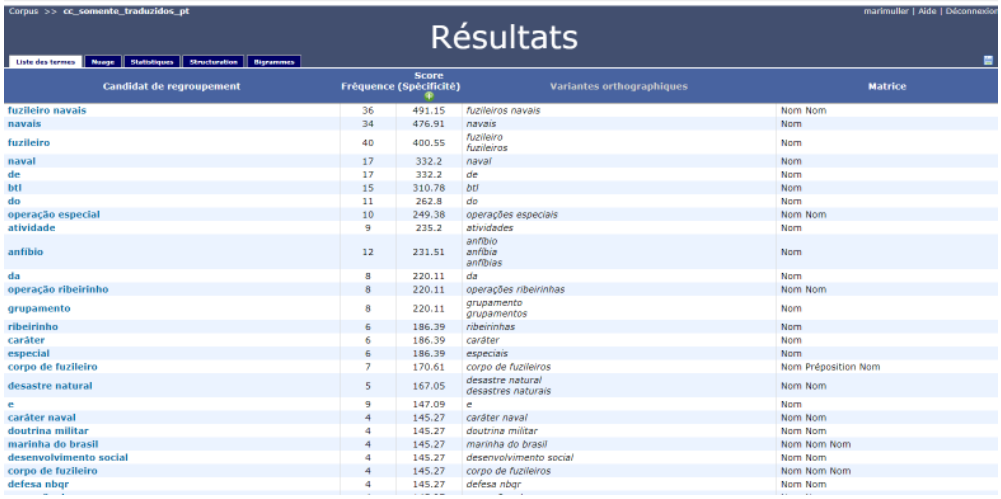


Figura 11 – Configurações selecionadas no TermoStat

Assim, o corpus de referência foi selecionado automaticamente a partir da escolha do idioma. Tanto para o português como para o inglês, o corpus é de domínio jornalístico. Assim, o próximo passo é iniciar a análise ao clicar em *Lancer l'analyse*, ou *Iniciar a análise*. Os resultados podem ser visualizados no formato de lista ou em nuvem. Foi escolhida a opção de ordená-los por chavicidade, a fim de manter a configuração escolhida para a EAT com o AntConc, mas também há a opção de organizar os termos por outras medidas de associação, como *CHI2*, *Log Likelihood* e *Log Odds Ratio*, o que gerou resultados bastante semelhantes com o

corpus de estudo utilizado nesta etapa, sem alterações significativas nos resultados. A imagem a seguir traz a tela que resultou na busca automática pelas palavras-chave e as informações que acompanharam os termos extraídos.



Candidat de regroupement	Score Fréquence (Spécificité)		Variantes orthographiques	Matrice
fuzileiro navais	36	491.15	*fuzileiros navais*	Nom Nom
navais	34	476.91	*navais*	Nom
fuzileiro	40	400.55	*fuzileiro* *fuzileiros*	Nom
naval	17	332.2	*naval*	Nom
de	17	332.2	*de*	Nom
btl	15	310.78	*btl*	Nom
do	11	262.8	*do*	Nom
operação especial	10	249.38	*operações especiais*	Nom Nom
atividade	9	235.2	*atividades*	Nom
anfíbio	12	231.51	*anfíbio* *anfíbia* *anfíbias*	Nom
da	8	220.11	*da*	Nom
operação ribeirinho	8	220.11	*operações ribeirinhas*	Nom Nom
grupamento	8	220.11	*grupamento* *grupamentos*	Nom
ribeirinho	6	186.39	*ribeirinhas*	Nom
caráter	6	186.39	*caráter*	Nom
especial	6	186.39	*especiais*	Nom
corpo de fuzileiro	7	170.61	*corpo de fuzileiros*	Nom Préposition Nom
desastre natural	5	167.05	*desastre natural* *desastres naturais*	Nom Nom
e	9	147.09	*e*	Nom
caráter naval	4	145.27	*caráter naval*	Nom Nom
doutrina militar	4	145.27	*doutrina militar*	Nom Nom
marinha do brasil	4	145.27	*marinha do brasil*	Nom Nom Nom
desenvolvimento social	4	145.27	*desenvolvimento social*	Nom Nom
corpo de fuzileiro	4	145.27	*corpo de fuzileiros*	Nom Nom Nom
defesa nbqr	4	145.27	*defesa nbqr*	Nom Nom

Figura 12 – Resultados da EAT do TermoStat

Foram selecionados pela ferramenta nessa fase 334 candidatos a termos por chavicidade para o português e apenas 62 candidatos para o inglês. As nuvens de termos extraídos geradas nos dois idiomas estão na figura 13.

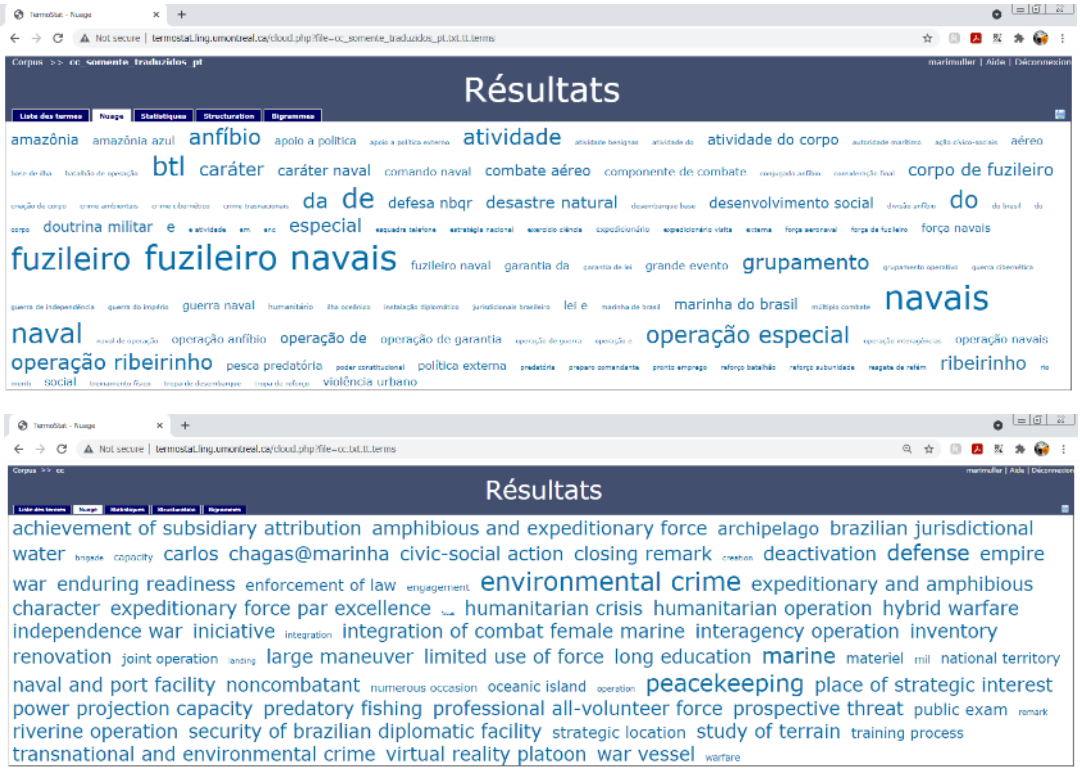


Figura 13 – Nuvens de palavras-chaves extraídas em português e em inglês

Pela análise das nuvens de palavras, pode-se perceber que a palavra-chave *fuzileiros navais* foi a de maior chavicidade do subcorpus em português, enquanto *environmental crime* foi a palavra-chave de maior chavicidade do subcorpus em inglês.

Por meio da análise das palavras-chave, o TermoStat também permite analisar as KWIC (*Key Word in Context*, ou palavra-chave em contexto). Aqui, clicamos na palavra *navais*, e em seguida na opção KWIC, e a seguinte tela foi mostrada.

Figura 14 – Resultado da opção KWIC para o termo *navais*

Ou seja, ainda que o TermoStat se proponha a entregar uma lista de n-gramas pronta, o programa também permite uma análise mais profunda das palavras-chave dentro do contexto. Na aba *Bigrams*, a ferramenta apresentou apenas a combinação *fuzileiros navais*, na qual classifica erroneamente *navais* como um verbo.

Os resultados da EAT do TermoStat foram separados em termos simples e complexos, eliminando aqueles compostos por mais de 4 itens lexicais[20]. O resultado foi o seguinte:

20. Foram mantidos apenas os termos compostos por até quatro itens lexicais por terem sido a maioria dos n-gramas identificados, sendo suficientes nesta etapa para fornecer dados referentes à eficiência das ferramentas na busca de termos complexos.

PALAVRAS-CHAVE EXTRAÍDAS PELO TERMOSTAT					
Português			Inglês		
Unigramas	Bigramas, trigramas e quadrigramas	**TOTAL**	Unigramas	Bigramas, trigramas e quadrigramas	**TOTAL**
93	224	317	22	39	61

Tabela 2 – Palavras-chave resultantes da EAT com o TermoStat

Convém apontar que o TermoStat pode omitir e modificar palavras gramaticais (*Marinha de Brasil*, em vez de *Marinha do Brasil*) e omitir plurais (*fuzileiro navais* em vez de *fuzileiros navais*), por isso a lista de palavras-chave fornecida pela ferramenta deve ser revisada e corrigida.

4.2.2
AVALIAÇÃO DO ANTCONC 3.5.7 COM LINHA DE CORTE

Primeiramente, após feito o upload do subcorpus de estudo em português no AntConc, foram selecionadas as preferências para a EAT na aba *Tool Preferences*. Quanto às medidas de chavicidade, independente da seleção realizada, devido ao tamanho reduzido do corpus utilizado no estudo piloto, não foram causadas alterações na lista de palavras-chave extraídas; logo, foram selecionados os primeiros itens de cada lista de valores (*Keyword Statistic*: *Log-Likelihood 4-term*; *Keyword Effect Size Measure*: *Dice Coefficient*).

Foi observado que, no TermoStat, o valor mínimo de chavicidade para uma palavra ser considerada chave foi de +5.67. Por isso, para o teste com o AntConc, foi selecionado um ponto de corte acima dessa medida. Assim, foi aplicada a medida de corte automático acima do valor mínimo considerado pelo TermoStat, ou p <0.01 (6.63). Logo, em *Keyword Effect Threshold*, foi selecionada essa opção, e em e *Keyword Effect Size Threshold*, foi selecionada a opção *All Values*, pois queríamos extrair todas as palavras-chave com o valor de chavicidade acima do fornecido como limite para o corte. Também experimentamos avaliar o AntConc sem a medida de corte, o que será descrito ao final desta seção.

Ainda em *Tool Preferences*, é possível inserir um ou mais arquivos para a função de corpus de referência, do acervo do usuário[21]. A figura a seguir apresenta a aba *Tool Preferences*, em que foram realizados os ajustes anteriormente mencionados.

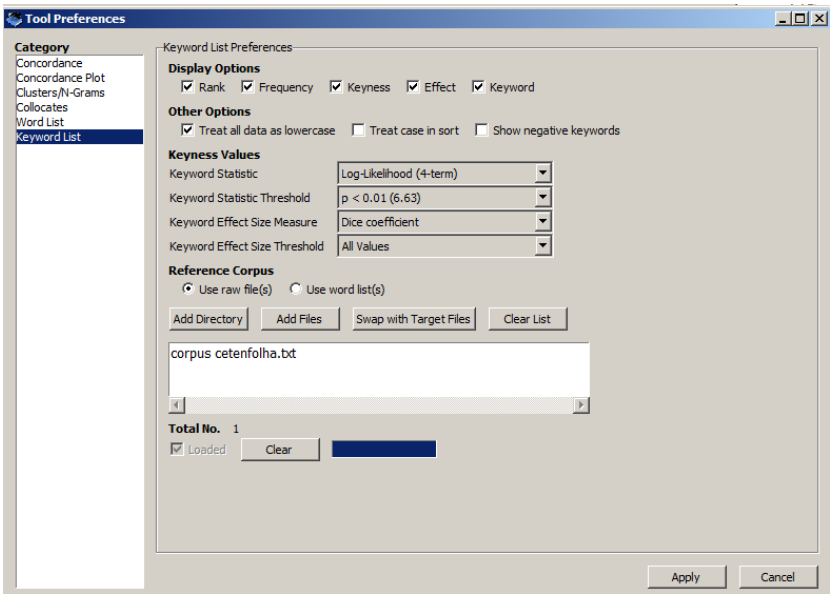


Figura 15 – Configurações selecionadas no AntConc (corpus de referência e medidas)

Na aba *Word List*, é possível inserir uma *stoplist*, a fim de eliminar palavras irrelevantes, como artigos e preposições, uma vez que o AntConc extrai automaticamente apenas unigramas. Logo, tais palavras, dada à alta frequência em um texto, poderiam ser consideradas termos se não fossem eliminadas. Aqui, foi inserida uma *stoplist* em português do acervo da autora, que contém palavras gramaticais.

Ajustadas as configurações, foi possível seguir para a EAT na aba *KeyWord List*, cujo resultado foi de 377 palavras-chave em português. Um extrato da seleção de palavras-chave organizadas por chavicidade e seus respectivos índices podem ser vistos na imagem a seguir.

21. Aqui, foi utilizado um extrato do corpus CETENFolha (2018), conforme consta na seção 3.1.3.

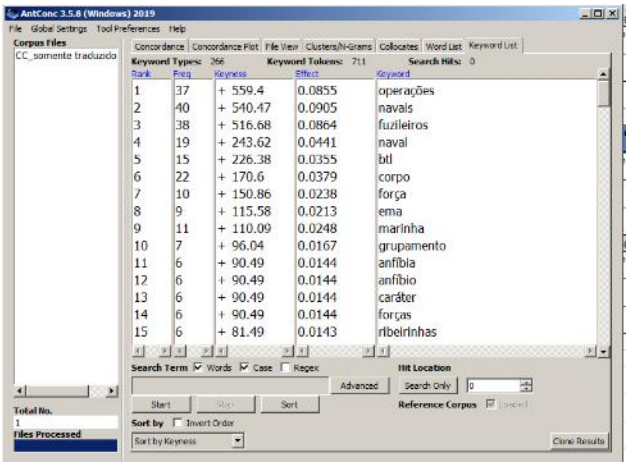


Figura 16 – Extrato da lista de palavras-chave em português gerada pelo AntConc

Na figura 16, é possível ver as 15 itens palavras-chave de maior chavicidade. A lista completa gerada foi selecionada, copiada e colada em planilha do Excel.

Ao utilizar a opção Clusters/N-Grams para buscar termos complexos, após a realização de diversos testes com configurações diferentes, foi escolhido o seguinte procedimento: ao clicar na opção *Advanced Search*, inserimos manualmente a lista de palavras-chave (unigramas) obtida na etapa anterior. A figura a seguir representa a aparência da janela do AntConc nesse ponto da pesquisa[22].

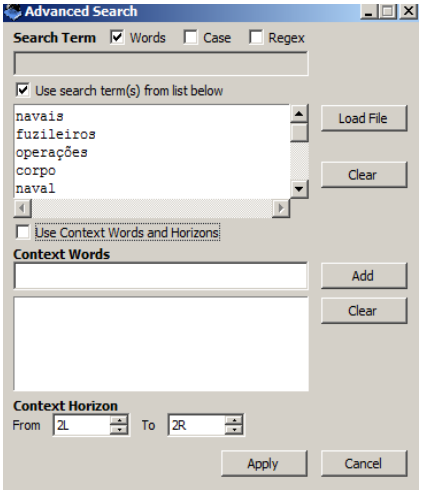


Figura 17 – Configuração da busca por n-gramas na opção clusters/n-grams

22. A configuração de *Context Horizon* foi mantida pois não interfere nesta etapa.

Aplicada a função anterior, o programa volta à tela de busca. Em seguida, foi selecionada, na opção *Cluster Size*, o número mínimo de 2 e máximo de 4 elementos, uma vez que buscamos n-gramas de até 4 itens lexicais. A frequência mínima foi configurada para 2, uma vez que a linha de corte de p <0.01 (6.63) não atinge a busca por termos complexos no programa.[23] A posição da palavra-chave selecionada foi à esquerda, devido à estrutura do português, em que os termos complexos crescem para a direita, isto é, o núcleo do sintagma nominal está à esquerda dos modificadores.

A janela *Collocates* (ou colocados) lista as palavras pela frequência com que aparecem nos contextos em relação ao termo pesquisado. A lista também indica a frequência com que as colocações ocorrem à esquerda ou à direita da palavra-chave. Assim, é uma ferramenta que ajuda a localizar os termos compostos e complexos, mas não fornece os termos prontos: é preciso analisar as linhas de concordância para validar ou não o candidato a termo.

A seguir, podemos ver o resultado de colocados ao fornecer a palavra-chave *navais* como nódulo de busca. A seleção foi organizada por frequência (selecionada a função *Sort by Freq* no canto esquerdo da tela). Em *Word Span*, foi mantida a seleção padrão de 5 palavras à esquerda e 5 à direita, dentro das quais os resultados deveriam estar.

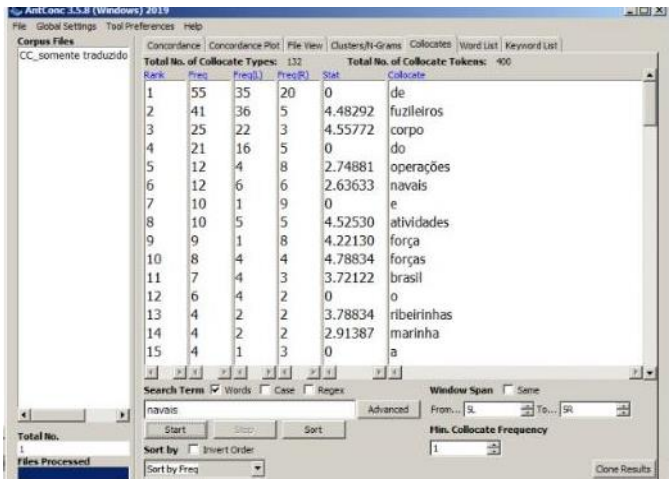


Figura 18 – Lista de colocados para a palavra-chave *navais* gerada pelo AntConc

23. Devido ao tamanho reduzido deste corpus, é esperado que se obtenha uma lista pequena de termos ao se optar por essa frequência, uma vez que a maioria dos n-gramas da LR aparecem apenas uma vez no corpus. Porém, em um corpus maior, esse efeito será reduzido, devido à repetição de termos proporcionalmente maior.

Como exemplo, clicamos na palavra *fuzileiros*, segunda colocada da lista em frequência, e a tela seguinte exibiu as linhas de concordância em que a palavra está inserida, conforme abaixo. Assim, pela análise de KWIC, ilustrada na figura a seguir, pôde-se confirmar que fuzileiros navais é realmente um termo.

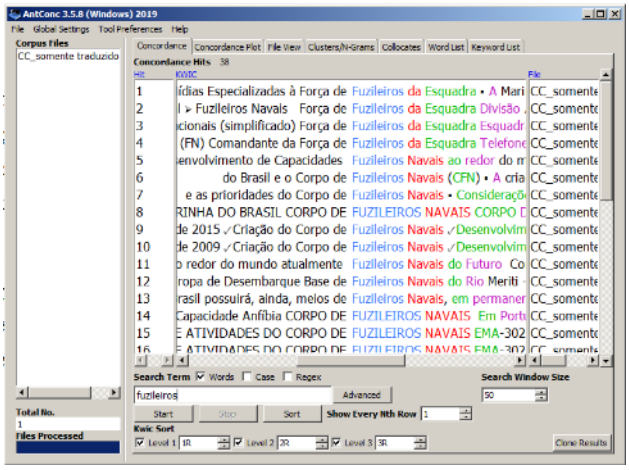


Figura 19 – Análise de KWIC da palavra-chave fuzileiros no AntConc

Depois disso, foi a vez de realizar os mesmos procedimentos com o subcorpus em inglês, inserindo a respectiva *stoplist*, também do acervo da autora, e o corpus de referência[24]. Assim, foram extraídas 64 palavras-chave (unigramas) em inglês.

Quanto à extração de bi/tri/quadrigramas, no inglês, a configuração que gerou o melhor resultado foi com o termo à esquerda, apesar de a posição à direita trazer termos diferentes e também verdadeiros positivos. A imagem a seguir mostra essa configuração.

24. Para o inglês, foi utilizado um extrato do corpus da revista TIME (s.d.), conforme consta na seção 3.1.3.

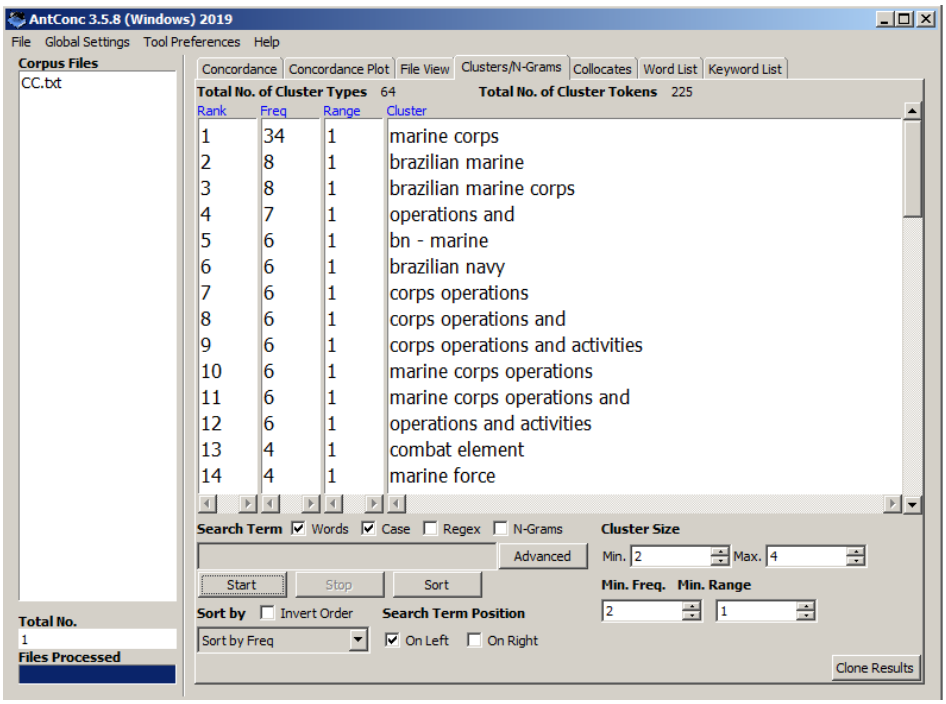


Figura 20 – Configuração da busca por n-gramas em inglês na opção clusters/n-grams

Essa lista foi copiada e inserida em uma planilha do Excel, ao lado de uma coluna onde já haviam sido relacionados os n-gramas de até quatro itens lexicais na LR.

Ao fim do processo, foram extraídas as seguintes quantidades de palavras-chaves com o AntConc, descritas na tabela 3.

PALAVRAS-CHAVE DO ANTCONC					
Português			Inglês		
Unigramas	Bi/tri/quadrigramas	**TOTAL**	Unigramas	Bi/tri/quadrigramas	**TOTAL**
266	144	410	45	64	109

Tabela 3 - Total de palavras-chave extraídas pelo AntConc

Após, a lista completa extraída foi comparada à LR e, assim, foram detectados os termos em comum nas duas listas, ou verdadeiros positivos. Depois da detecção dos verdadeiros positivos, foram calculadas a precisão, a abrangência e a medida F para os candidatos extraídos.

4.2.3
AVALIAÇÃO DO ANTCONC SEM LINHA DE CORTE

A fim de testar todo o potencial do AntConc sem limitar suas configurações, conforme realizado ao procurar aproximá-las das configurações automáticas do TermoStat para comparar a eficácia das duas ferramentas, foi feito um teste de utilização híbrida da ferramenta, com a análise manual das linhas de concordância de cada palavra-chave fornecida na pesquisa. Ao utilizar o AntConc sem o corte, com o mesmo corpus de referência, e realizando a análise manual das linhas de concordância das palavras-chaves fornecidas, os seguintes resultados foram obtidos:

Idioma	Ocorrências	Unigramas	Bi/Tri/ Quadrigramas	PRECISÃO	ABRANGÊNCIA	MEDIDA F
Português	395	50	98	37%	100%	54%
Inglês	379	51	107	41%	100%	58%

Tabela 4 – Resultados dos testes de análise híbrida com o AntConc

Assim, observa-se que foi obtido 100% dos termos da LR, tanto em português quanto em inglês.

A seguir, serão comparados os resultados dos testes com as duas ferramentas a fim de identificar quais tiveram o melhor desempenho para unigramas, bi/tri/quadrigramas e para os termos extraídos em geral.

4.3
Resultados e análise

No total, os testes com as ferramentas com linha de corte resultaram na extração da seguinte quantidade de palavras-chave:

PALAVRAS-CHAVE DO TERMOSTAT					
Português			Inglês		
Unigra-mas	Bi/tri/ quadrigramas	**TOTAL**	Unigra-mas	Bi/tri/ quadrigramas	**TOTAL**
93	224	**317**	22	39	**61**
PALAVRAS-CHAVE DO ANTCONC					
Português			Inglês		
Unigra-mas	Bi/tri/ quadrigramas	**TOTAL**	Unigra-mas	Bi/tri/ quadrigramas	**TOTAL**
266	144	**410**	45	64	**109**

Tabela 5 - Comparação da quantidade de palavras-chave extraídas pelas ferramentas

As listas de termos extraídos de ambas as ferramentas que foram validados podem ser vistas na íntegra nos Apêndices 5 a 8. Após o contraste da lista extraída com a LR, obtivemos os verdadeiros positivos em português. Essa comparação gerou os seguintes resultados:

	Unigramas				
Ferramenta	Ocorrências	Verdadeiros positivos	Precisão	Abrangência	Medida F
AntConc	266	39	3%	78%	24%
TermoStat	93	28	30%	56%	39%
	Bi/Tri/Quadrigramas				
Ferramenta	Ocorrências	Verdadeiros positivos	Precisão	Abrangência	Medida F
AntConc	144	16	11%	16%	13%
TermoStat	224	30	13%	30%	18%

Tabela 6 – Resultados dos testes com o subcorpus em português

Observa-se que, ao analisar detalhadamente a tabela dos testes em português, o TermoStat (medida F= 39% para unigramas e 18% para bi a quadrigramas) apresentou resultados melhores do que o AntConc (medida F= 24% para unigramas e 13% para bigramas a quadrigramas). Tal diferença se deu principalmente devido ao TermoStat apresentar uma precisão muito maior para unigramas (30% contra 3%

do AntConc). Logo, o TermoStat obteve as melhores medidas F tanto para a extração de unigramas quanto de termos compostos/complexos até quadrigramas.

Já a análise do material em inglês gerou os seguintes resultados:

	Unigramas				
Ferramenta	Ocorrências	Verdadeiros positivos	Precisão	Abrangência	Medida F
AntConc	45	15	33%	29%	31%
TermoStat	16	5	31%	9%	14%
	Bi/Tri/Quadrigramas				
Ferramenta	Ocorrências	Verdadeiros positivos	Precisão	Abrangência	Medida F
AntConc	64	8	10%	17%	12%
TermoStat	39	10	25%	9%	13%

Tabela 7 – Resultados dos testes com o subcorpus em inglês

Conforme observado, em relação ao corpus em inglês, o AntConc se destacou na extração de unigramas, por apresentar precisão e abrangência maiores, com a medida F de 31% contra 14% do TermoStat. Já em relação a termos complexos, apesar de as duas ferramentas apresentarem resultados muito próximos, o TermoStat continuou se destacando (medida F= 13%) em relação ao AntConc (medida F= 12%).

Comparando o desempenho das ferramentas em relação ao idioma analisado, pudemos obter os seguintes resultados, resumidamente:

	Português	**Inglês**
AntConc	410	109
TermoStat	317	61

Tabela 8 – Quantidade de candidatos a termos extraídos por idioma

Observa-se que a lista de EAT em inglês gerada por ambas as ferramentas foi muito menor do que a em português no caso dos dois idiomas (3,76 vezes maior no AntConc e 5,19 vezes no TermoStat). Tal fato se justifica pelo uso do ponto de corte utilizado, que considerou itens lexicais de no mínimo + 6.79 de chavicidade. Um resultado semelhante foi obtido no AntConc ao limitar a busca com itens acima

da 6.63 de chavicidade ($p < 0.01$). Um estudo posterior da ocorrência de termos militares em corpora de textos jornalísticos em inglês seria necessário para entender o fenômeno ocorrido.

Para testar esse fenômeno, foi feita a busca do termo de maior chavicidade do português no AntConc (*fuzileiro naval*) em um extrato de 50% do corpus de referência em português. O resultado da busca está ilustrado na figura a seguir.

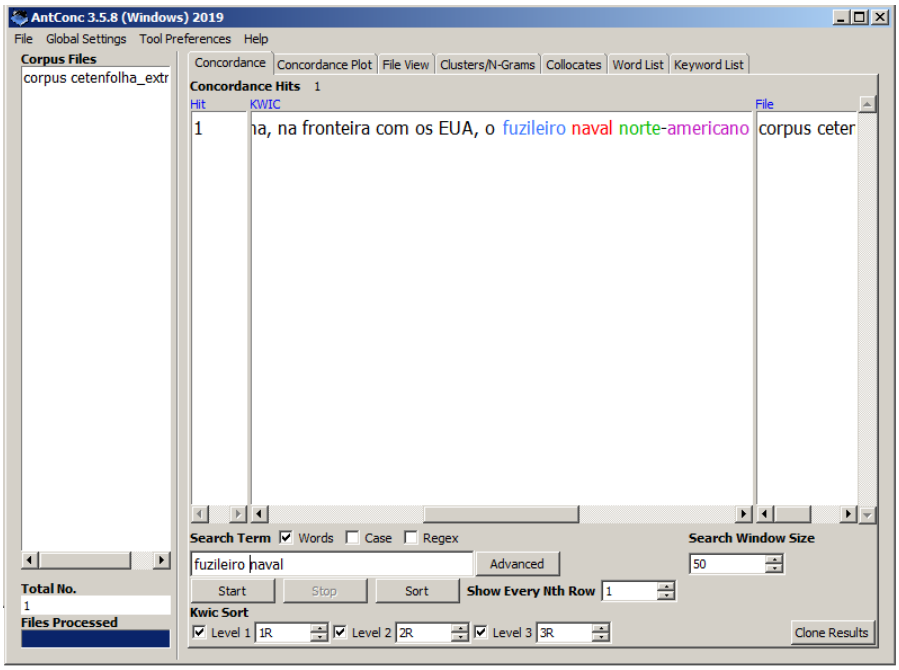


Figura 21 – Resultado da busca pelo termo *fuzileiro naval* no extrato do corpus de referência em português

Da mesma forma, foi feita a busca do termo de maior chavicidade do inglês (*marine*) em um extrato de 50% do corpus de referência em inglês, conforme a figura a seguir.

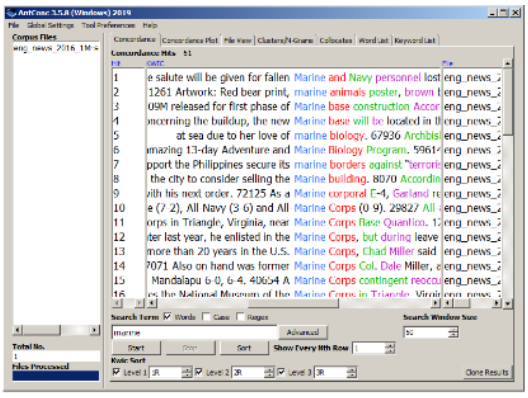


Figura 22– Resultado da busca pelo termo *marine* no extrato do corpus de referência em inglês

Tais amostras foram limitadas a 50% dos corpora de referência pois a ferramenta trava ao tentar analisar um corpus muito grande. Foi observado que houve apenas uma ocorrência desse termo no extrato de corpus de referência em português (figura 21) e 61 ocorrências no extrato do corpus de referência em inglês (figura 22). Ou seja, considerando a amostra, o termo é 6.000% mais frequente nos textos jornalísticos em inglês do que nos textos jornalísticos em português.

Já a tabela 9 permite comparar os resultados obtidos por meio da análise híbrida com o auxílio do AntConc com os resultados dos testes resultantes das análises automáticas realizadas com as duas ferramentas, considerando o total de termos extraídos (de unigramas a quadrigramas).

ANÁLISE HÍBRIDA - ANTCONC			
IDIOMA	Precisão	Abrangência	Medida F
Português	37%	100%	54%
Inglês	41%	100%	58%
ANÁLISE AUTOMÁTICA - ANTCONC			
IDIOMA	Precisão	Abrangência	Medida F
Português	13%	37%	19%
Inglês	21%	14%	17%
ANÁLISE AUTOMÁTICA - TERMOSTAT			
IDIOMA	Precisão	Abrangência	Medida F
Português	18%	39%	24%
Inglês	24%	8%	13%

Tabela 9 – Resultados dos testes de análise híbrida e automática com o AntConc e da análise automática com o TermoStat nos dois idiomas de estudo

Apesar das medidas muito baixas das análises automáticas, vale lembrar que a EAT é apenas o primeiro passo de uma análise terminológica. Aqui, porém, a intenção foi de avaliar a lista extraída de forma totalmente automática, a fim de medir o trabalho humano necessário na elaboração de uma terminologia de qualidade.

No total, na análise automática, o TermoStat se destacou na detecção de termos em português (medida F 24% contra medida F 19% do AntConc) e o AntConc se destacou na detecção de termos em inglês (medida F 17% contra medida F 13% do TermoStat).

Porém, é possível constatar o aumento significativo de todos os índices com a extração híbrida. Comparando os índices das medidas F, temos: 54% x 19% (AntConc híbrido x AntConc automático em português); 58% x 17% (AntConc híbrido x AntConc automático em inglês); 54% x 24% (AntConc híbrido x TermoStat automático em português); 58% x 13% (AntConc híbrido x TermoStat automático em inglês).

Após a análise de todos os testes realizados com as ferramentas, os resultados sugerem que, caso haja tempo para realizar manualmente a análise das KWIC, que utilize preferencialmente o AntConc sem a linha de corte, analisando as palavras-chaves em contexto. Já para o caso de ser necessário realizar uma extração de termos mais rápida, sugere-se o uso do TermoStat, devido aos melhores índices aqui apresentados.

Neste estudo, a fim de obter os melhores resultados possíveis (precisos e abrangentes) da subárea do CFN, optou-se por realizar o uso do AntConc sem o corte. O recurso de utilizar a linha de corte seria interessante, conforme exposto, para uma busca mais rápida, devido à grande quantidade de termos que se concentram dentro desse limite, visto que os termos mais específicos e frequentes encabeçam as listas. Assim, o uso do corte facilitaria a EAT desse corpus, cerca de 80 vezes maior que o utilizado no estudo piloto, que gerará uma lista de cerca de 28.000 candidatos a termos. Porém, convém lembrar que essa opção não elimina a necessidade da análise das linhas de concordância dos itens que não estiverem dentro dessa seleção, a fim de resgatar possíveis verdadeiros positivos que não apre-

sentem ocorrência muito frequente no corpus, mas que ainda assim podem ser termos relevantes. Portanto, optou-se por realizar um estudo mais detalhado dos resultados fornecidos pelo AntConc.

Com base nos resultados apresentados, conclui-se que, para a extração dos termos do corpus em estudo, o TermoStat se saiu melhor nos testes apenas automáticos, sem análise humana. Porém, ainda que a utilização do TermoStat seja mais prática por este já apresentar os termos complexos listados, necessitando, apenas, de revisão e possível correção. a utilização do AntConc é mais trabalhosa, pois necessita de diversos ajustes para sua utilização. Porém, sua análise gera resultados mais precisos e alinhados com os objetivos do analista, com abrangência de 100% ao analisar manualmente as palavras-chaves nas linhas de concordância.

Outro fator interessante é que o AntConc permite que o usuário escolha aplicar ou não a linha de corte, e, inclusive, ajustar essa linha de acordo com o seu objetivo. É importante lembrar que, neste estudo, os termos que comporão as entradas do léxico serão extraídos do corpus em português, sendo o corpus em inglês utilizado apenas para extrair termos equivalentes. Por isso, para a detecção desses equivalentes terminológicos em inglês, será utilizado, preferencialmente, o AntConc.

Por fim, convém ressaltar que a Lista de Referência gerada no estudo piloto também será utilizada para compor o léxico final deste estudo, após a seleção de termos do CFN entre os termos militares genéricos.

A fim de se estabelecer uma comparação com outros estudos, vale lembrar que nos testes de Teixeira (2010), que realizou um estudo comparando quatro ferramentas de EAT: Corpógrafo 4.0, WordSmith Tools 3.0, e-Termos e ZExtractor; o Corpógrafo foi o que apresentou melhor resultado, com 27.56% de acerto (precisão) na extração de unigramas. Nesses testes, foram inseridos linhas de cortes, além de realizada a análise manual de linhas de concordância, para se detectar os termos.

5
O LÉXICO *ESPÍRITO DE CORPUS*

Neste capítulo, cobriremos o processo de extração de termos do corpus de estudo para compor o léxico *Espírito de Corpus*, conforme a configuração estabelecida após a realização dos testes com o estudo piloto, descritos no capítulo 4. Optou-se, dessa forma, por realizar a extração de termos de forma híbrida (automática e manual), com a utilização da ferramenta AntConc, sem utilizar linha de corte. Os termos identificados no estudo piloto também serão inseridos no léxico final, eliminando os itens repetidos que aparecerem nessa nova etapa da busca por termos. A análise manual dos termos seguirá os critérios da Semântica Lexical propostos por L'Homme (2020).

Convém lembrar que no estudo piloto, não foram utilizados esses critérios para a validação dos termos, mas apenas a opinião da especialista e a comparação dos itens das LR com os de outras listas de referência, uma vez que, ali, o objetivo era apenas de verificar o desempenho das ferramentas.

Uma vez que o processo de extração automática já foi descrito detalhadamente no capítulo anterior, o foco aqui será em descrever a análise manual dos termos e as demais observações julgadas pertinentes. Já os procedimentos automáticos e o processo de publicação do léxico em si na internet serão mencionados brevemente.

A intenção da criação do léxico bilíngue do CFN não foi de esgotar os termos da subárea em questão, mas sim de focar em termos identificados como problemáticos no âmbito da tradução desse domínio, como patentes, nomes de Organizações Militares, incluindo também funções e definições específicas do CFN. Além disso, procurou-se oferecer equivalentes terminológicos para termos militares comuns no contexto do CFN, a fim de fornecer um material confiável de pesquisa àqueles que necessitem realizar traduções de textos do Corpo de Fuzileiros Navais do português para o inglês, mas também úteis para quem desejar realizar versões de textos do USMC do inglês para o português.

Apesar de já haver no mercado material terminológico monolíngue e bilíngue especializado na área militar (DEPARTMENT OF DEFENSE, 2007; U.S.

ARMY, 2015) considerados confiáveis e de referência, a intenção aqui foi de utilizar também esse material para complementar o léxico criado por meio da análise do corpus de estudo e, assim, unir a esse produto terminológico informações relevantes para a subárea do CFN.

O processo de compilação dos termos para o léxico contou com as seguintes etapas e subetapas, evidenciadas na figura 23[25]:

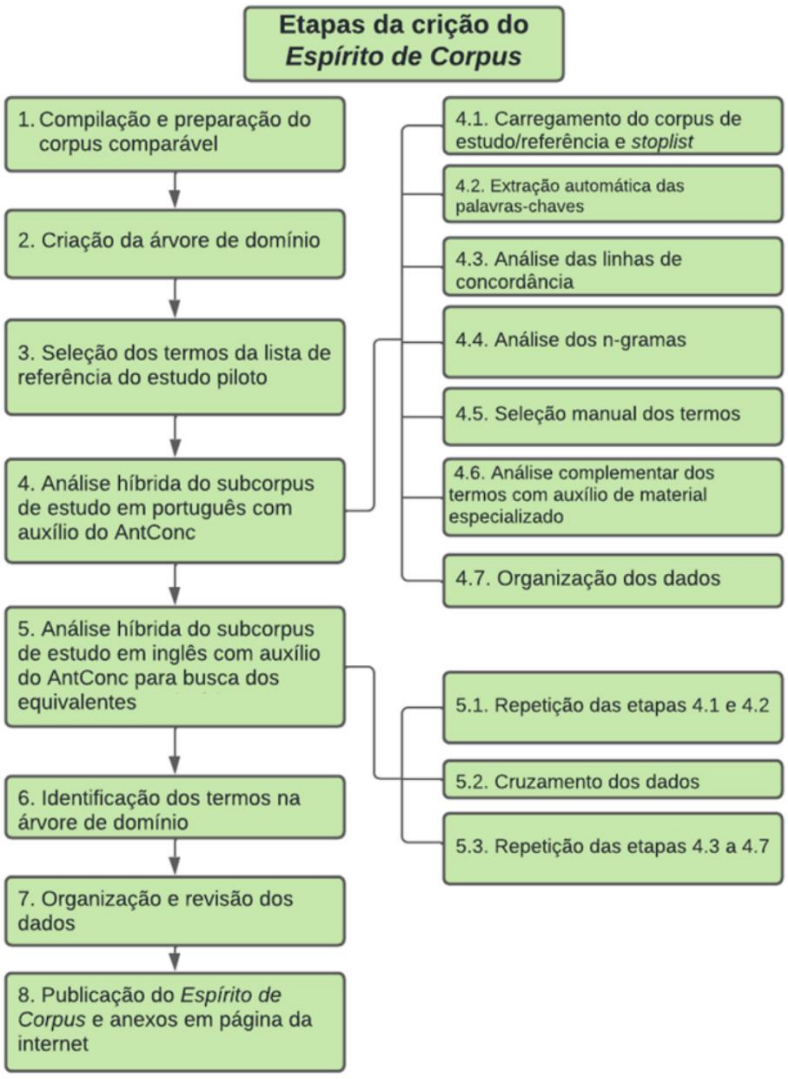


Figura 23 – Etapas da criação do *Espírito de Corpus*

[25] A figura 23 é um extrato da figura 7, apresentada no capítulo 3.

5.1
CRIAÇÃO DA ÁRVORE DE DOMÍNIO

Conforme mencionado no capítulo 2.1, a árvore de domínio é o ponto de partida para a estruturação da área do conhecimento em que o estudo terminológico se baseia, permitindo organizar o conhecimento. Também, conforme mencionado, a árvore de domínio da qual partimos neste estudo foi a utilizada em Müller (2019).

Neste trabalho, devido às suas peculiaridades e para uma melhor organização dos termos abrangidos pelo léxico, foi necessário modificar um pouco a estrutura da árvore de Müller (2019) em função dos dados do corpus, uma vez que pretendemos realizar um estudo muito mais detalhado e abrangente. Foram mantidas apenas as classes que seriam cobertas pelo estudo. A caixa *Componentes*, subordinada a *Doutrina*, foi excluída. Foi criado o campo *Organizações Militares*, que ganhou destaque uma vez que se pretendeu, neste estudo, dar maior atenção a esses termos. Já o campo *Pessoal* foi dividido em *Patentes* e em *Funções*. Optou-se por excluir a caixa *Projetos*, por não serem cobertos pelo corpus de estudo, e em seu lugar foi adicionada a caixa *Programas*. O resultado pode ser visto na figura a seguir.

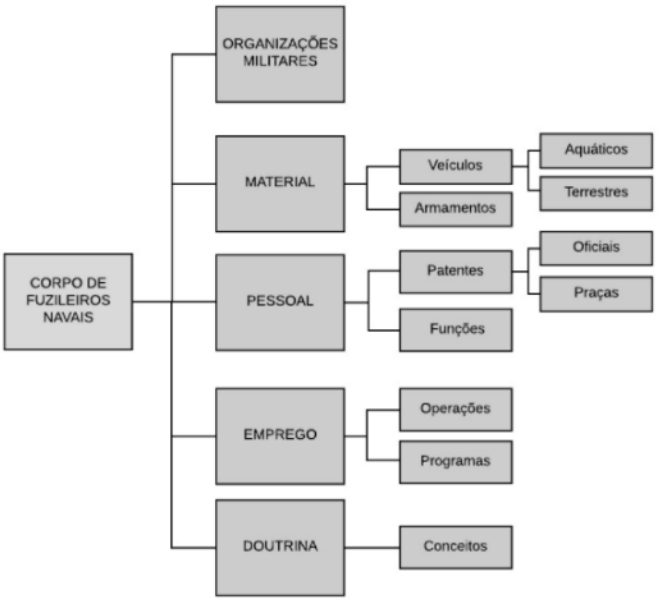


Figura 24 – Árvore de domínio atualizada

Conforme pode ser visto na figura 24, diferentemente da árvore de domínio de Müller (2019), esta possui apenas as classes abrangidas pelo corpus e não os termos a elas pertencentes, dada a quantidade de termos (271).

5.2
SELEÇÃO DOS TERMOS DA LISTA DE REFERÊNCIA DO ESTUDO PILOTO

Primeiramente, foram analisadas as listas de referência de termos militares fruto dos testes realizados no estudo piloto. Após a análise da especialista, foi decidido, a princípio, manter toda a lista, uma vez que, ainda que possua termos genéricos da área militar, como *ataque*, *combate* e *tropa*, esses são termos relevantes para o CFN[26]. Essa lista se encontra na íntegra nos apêndices1 e 2.

5.3
ANÁLISE HÍBRIDA DO CORPUS DE ESTUDO EM PORTUGUÊS COM AUXÍLIO DO ANTCONC

Esta fase seguiu as seguintes etapas, já mencionadas na figura 23. Iniciou-se a análise do corpus de estudo na ferramenta AntConc por meio do carregamento do corpus de estudo em português (etapa 4.1), composto por textos do CFN, e do respectivo corpus de referência, composto por textos jornalísticos. Em seguida, foi feito o carregamento da lista de *Stop Words*, composto por palavras gramaticais (a mesma utilizada no estudo piloto).

Foi feita a extração dos unigramas com e sem a linha de corte (etapa 4.2). Essas listas foram tratadas separadamente. Conforme estabelecido no capítulo anterior, não foi utilizada qualquer linha de corte. Assim, a extração de palavras-chaves sem linha de corte gerou uma lista de 7.626 entradas.

Para configuração da ferramenta na busca de termos complexos (etapa 4.3), foi selecionada, na opção *Clusters/N-grams>Cluster Size*,o número mínimo de 2 e máximo de 8 elementos; a frequência mínima foi configurada para 2. A posição do

26. Após a extração automática de termos, foi observado que alguns termos genéricos da área militar apareciam com muita frequência no corpus de estudo (*ataque*: chavicidade de +404.66, 169 aparições no corpus; *combate*: chavicidade de + 275.69, 104 aparições; *tropa*: chavicidade de +536.99, 124 aparições).

termo foi buscada tanto à direita quanto à esquerda. A lista de *Clusters* gerou 20.301 entradas.

Todos os candidatos a termos (CT) extraídos automaticamente pelo AntConc nesta etapa do estudo foram adicionados à tabela dos termos já validados no estudo piloto.[27] Os candidatos a termos que já haviam sido detectados pelo estudo piloto e estavam repetidos foram, primeiramente, eliminados. Em seguida, a lista restante foi submetida à análise manual das linhas de concordância sob os critérios de análise da Semântica Lexical segundo L'Homme (2020), conforme descritas no capítulo 2.3.3. Então, ao lado da coluna *Termos_Pt*, foram inseridas quatro colunas para indicação dos critérios em que os candidatos a termos se enquadravam. Um extrato dessa parte da tabela será mostrado no quadro 9 da seção 5.3.

Em relação ao critério A, foi feita uma pré-seleção da lista fornecida automaticamente pela ferramenta, eliminando assim todos os candidatos a termos que, apesar de serem classificados pelo AntConc com chavicidade positiva, não tinham relação específica com o campo de conhecimento do CFN nem eram parte de termos compostos ou complexos desse campo de estudo; portanto, foram considerados pela autora como palavras sem caráter terminológico. Alguns desses candidatos a termos eliminados, que compunham a maioria dos itens da lista, foram *areia*, *cardíaco, consideração, cortesia, importância* e *possuir*.

Partiu-se então para a análise e seleção dos termos que permaneceram na lista em relação ao critério B, ou à natureza dos argumentos. Nesse ponto, foram selecionados nódulos frequentes que já haviam sido identificados como termos no âmbito do CFN (como os adjetivos *anfíbio* e *naval* e os substantivos *força* e *comando*) e, a partir deles, buscados termos compostos e complexos na aba *Clusters/N-grams*. Ou seja, foi realizada uma busca mais direcionada por n-gramas do que a realizada na etapa 4.3, com foco agora nos nódulos frequentes já identificados como termos no estudo piloto. Assim, foram analisadas as linhas de concordância e adicionados os n-gramas que continham tais palavras como nódulo[28]. A tabela a seguir exemplifica essa etapa da análise.

27. Essa tabela do Excel, nomeada de *Léxico pt_Antconc*, continha as seguintes colunas: *Termos_Pt*, *Variações_Pt*, *Subárea*, *Termos_En* e *Variações_En.*

28Alguns termos que já constavam na lista de referência do estudo piloto, como *Divisão Anfíbia*, *Comando da Tropa de Reforço* e *Batalhão Naval*, apareceram novamente nessa etapa da pesquisa.

anfíbio(a) (adj.)	assalto anfíbio carro lagarta anfíbio combatente anfíbio Divisão Anfíbia demonstração anfíbia
naval (adj.)	batalha naval artilharia naval Batalhão Naval Comandante do Batalhão Naval distrito naval
força (subst.)	força atacante força de cobertura força de desembarque força-tarefa anfíbia
comando (subst.)	Comando da Divisão Anfíbia (ComDivAnf) Comando da Força de Fuzileiros da Esquadra (ComFFE) Comando da Tropa de Reforço (ComTrRef) Comando de Operações Navais (ComOpNav)

Quadro 6 – N-gramas encontrados após busca a partir de nódulos frequentes (critério B da Semântica Lexical)

Depois, foi realizada a identificação de termos em relação ao critério C, que propõe a análise das relações morfológicas e semânticas existentes entre os termos. Tal critério permitiu validar e/ou incluir na lista termos com o mesmo radical, que formavam "famílias" de termos, conforme exposto na tabela seguinte.

Radical	Família de termos
local	*deslocamento*; *deslocar*
guarda	*vanguarda*; *retaguarda*; *flancoguarda;* *ponta de vanguarda*
comando	*comandante*
embarque	*embarcação*; *desembarque*
camuflado	*camuflagem*

Quadro 7 – Termos encontrados após busca a partir de relações morfológicas e semânticas identificadas (critério C da Semântica Lexical)

Também se buscou identificar termos a partir do critério D, referente às relações paradigmáticas existentes entre os termos. Este último critério auxiliou também na inserção dos termos na árvore de domínio. Algumas relações paradigmáticas identificadas podem ser observadas na tabela a seguir, referente a Patentes (inseridos na árvore de domínio no campo subordinado a Pessoal). Assim, observa-se que *Oficiais Generais* é hiperônimo de *Almirante*, enquanto *Oficiais Superiores* é hiperônimo de *Capitão de Mar e Guerra*. Já *Cabo*, *Soldado* e *Recruta* são hipônimos de *Praças não graduadas*. Algumas dessas patentes não existiam no corpus de estudo, mas foram adicionadas a fim de completar essa rede de relações paradigmáticas, uma vez que, segundo a L'Homme (2020), constituem termos. Dessa forma, foi possível apresentar no léxico uma versão completa dessa subárea, considerada crítica para a tradução. Assim, formou-se uma tabela completa de patentes de Oficiais e Praças, que mais tarde foi inserida no léxico com seus respectivos equivalentes terminológicos. A primeira versão dessa tabela, apenas em português, está representada a seguir.

<table>
<tr><th colspan="4">PESSOAL</th></tr>
<tr><th colspan="2">Oficiais</th><th colspan="2">Praças</th></tr>
<tr><td rowspan="4">Oficiais Generais</td><td>Almirante</td><td rowspan="6">Praças Graduadas</td><td rowspan="2">Suboficial-Mor</td></tr>
<tr><td>Almirante de Esquadra</td></tr>
<tr><td>Vice-Almirante</td><td>Suboficial</td></tr>
<tr><td>Contra-Almirante</td><td>Primeiro-Sargento</td></tr>
<tr><td rowspan="3">Oficiais Superiores</td><td>Capitão de Mar e Guerra</td><td>Segundo-Sargento</td></tr>
<tr><td>Capitão de Fragata</td><td>Terceiro-Sargento</td></tr>
<tr><td>Capitão de Corveta</td><td rowspan="4">Praças Não-Graduadas</td><td>Cabo</td></tr>
<tr><td rowspan="2">Oficiais Intermediários</td><td rowspan="2">Capitão-Tenente</td><td rowspan="2">Soldado</td></tr>
<tr><td rowspan="2">Oficiais Subalternos</td><td rowspan="2">Recruta</td></tr>
<tr><td>Primeiro-Tenente</td></tr>
<tr><td>Segundo-Tenente</td><td colspan="2"></td></tr>
<tr><td>Praças especiais</td><td>Guarda-Marinha</td><td colspan="2"></td></tr>
</table>

Quadro 8 – Organização em tabela dos termos referentes a patentes para análise das relações paradigmáticas (critério D da Semântica Lexical)

Da mesma forma, outras relações paradigmáticas puderam ser identificadas, como as *Organizações Militares* (OM) subordinadas ao *Comando de Operações Navais*; as OM subordinadas ao *Comando-Geral do CFN*, todas essas hiperônimos de CFN, que, por sua vez, é hipônimo de *Marinha do Brasil*. Quanto aos nomes das *Organizações Militares* (OM) do CFN, optou-se por inseri-los, além de na listagem do léxico, na forma de organograma, para melhor visualização quanto às relações de subordinação estabelecidas, com a utilização de siglas. A figura a seguir mostra esse organograma nesse ponto da pesquisa.

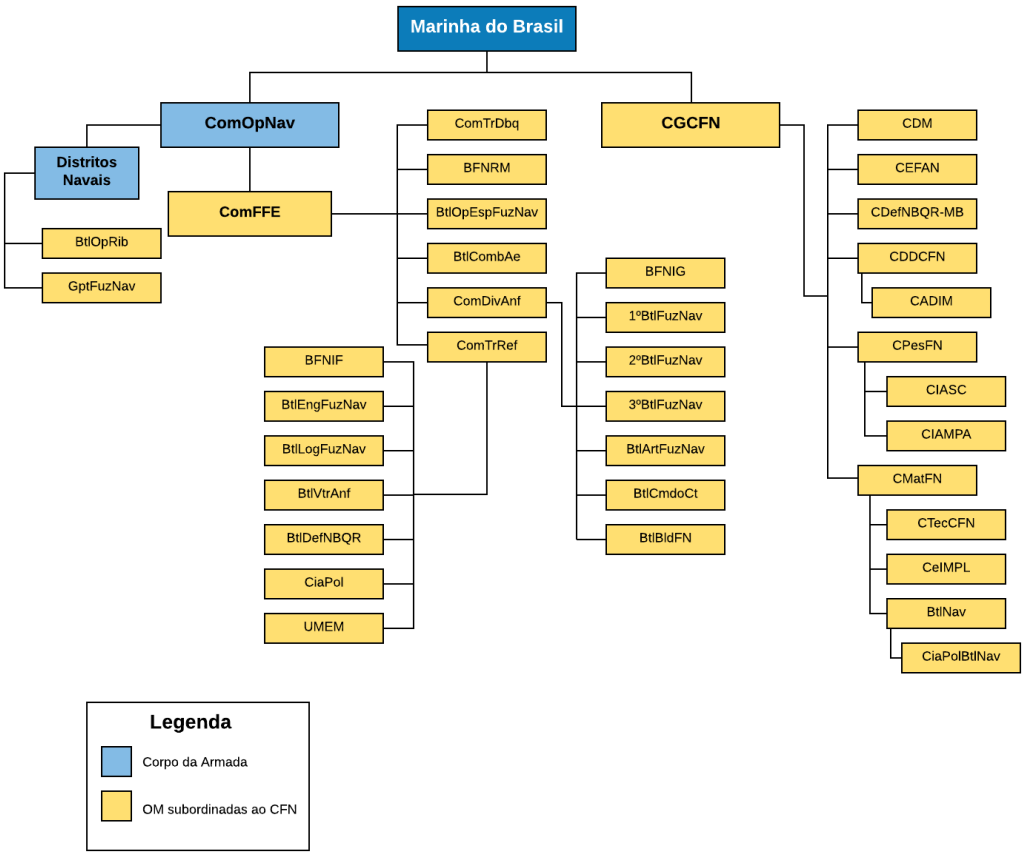


Figura 25 – Relações paradigmáticas entre as OM do CFN (critério D da Semântica Lexical)

Ao final da análise quanto aos quatro critérios da Semântica Lexical, foi indicado com um *X* em uma coluna ao lado dos termos listados na tabela do Excel os critérios que se aplicavam a eles. Segue um extrato dessa tabela.

TERMOS	CRITÉRIOS – SEMÂNTICA LEXICAL			
	A	B	C	D
cobertas	X			X
comandante	X	X	X	X
comando	X	X	X	X
combate	X	X	X	X
combatente	X	X	X	X
continência	X			X
controle	X	X		X
defesa	X	X		X
desembarque	X	X	X	X
deslocamento	X	X	X	X

Quadro 9 – Extrato da tabela de termos avaliados quanto aos critérios da Semântica Lexical

Conforme pode ser observado no quadro 11, alguns dos termos que cumpriram com todos os quatro critérios da Semântica Lexical foram *comandante, comando, combate, combatente, desembarque e deslocamento*. Já os termos *cobertas* e *continência* não cumpriram com os critérios em relação à natureza dos argumentos nem estabeleceram relações morfológicas e semânticas com outros termos identificados anteriormente. Porém, ainda assim, foram considerados termos, pois atenderam a pelo menos dois critérios, assim como todos os outros termos inseridos no léxico.

Embora no estudo piloto tenham sido inseridos, na lista de referência, apenas termos constantes no corpus de estudo, uma vez que o objetivo dessa etapa tenha sido de avaliar a precisão e a abrangência das ferramentas de EAT, na confecção do léxico, foram incluídos também termos não existentes no corpus, mas que foram identificados como pertencentes à mesma família semântica ou que estabeleciam relações paradigmáticas importantes entre termos constantes no material de estudo. O motivo para tal foi de enriquecer o material e de fornecer a maior quantidade possível de termos relevantes dentro de determinado subdomínio.

Quanto às siglas e abreviaturas em português, foram consultados, além do corpus em questão, o Manual de Abreviaturas, Siglas, Símbolos e Convenções Cartográficas das Forças Armadas (BRASIL, 2008). Esse manual foi utilizado para checar siglas e abreviaturas de termos listados no léxico que não existiam no corpus, como de algumas patentes militares (*Contra-Almirante*: *CA* ou *C Alte*, por exemplo), além de trazer informações complementares em relação àquelas siglas que só

são utilizadas em conjunto, como no caso de *guerra* (G) e *posto* (P). Essas informações também foram adicionadas ao léxico. Assim, a consulta a esse manual foi de suma importância para complementar as informações extraídas da análise do corpus.

5.4
ANÁLISE HÍBRIDA DO CORPUS DE ESTUDO EM INGLÊS COM AUXÍLIO DO ANTCONC PARA BUSCA DOS EQUIVALENTES TERMINOLÓGICOS

Após extraídos e validados os termos em português, foi a vez de buscar os equivalentes terminológicos no corpus em inglês. No total, foram listados no léxico 271 termos com variações e equivalentes terminológicos em inglês. Neste capítulo, vamos abordar os casos mais interessantes observados nesta etapa em relação à significação dos termos e sua equivalência em inglês.

De forma geral, para a busca dos equivalentes terminológicos, foram seguidas as etapas já mencionadas na figura 23. Assim como realizado com o corpus em português, foi feito o carregamento no AntConc do corpus de estudo e de referência em inglês, além da inserção na ferramenta da mesma *StopList* utilizada no estudo piloto (etapa 5.1). Depois, foi feita a extração automática das palavras-chaves em inglês (5.2), que também foram inseridos em uma tabela específica no Excel, nomeada como *Léxico En_Antconc*. A etapa de análise das linhas de concordância (5.4) foi realizada após o cruzamento dos dados (5.3), ou seja, após a associação entre os candidatos a termos extraídos automaticamente do corpus em inglês e os termos extraídos do corpus em português, para melhor entendimento do sentido dos termos, a fim de verificar se estabeleciam ou não uma relação de equivalência com os termos em português, com a análise dos termos em seu contexto de origem.

Conforme mencionado no capítulo 2.1.3, para tratar a equivalência entre termos de diferentes línguas extraídos de corpora, L'Homme (2004; 2020) sugere a utilização das abordagens léxico-semânticas, realizadas por meio da observação das relações que os itens lexicais estabelecem entre si no contexto de origem. A fim de evitar ambiguidades e dúvidas ao usuário do léxico nos casos de polissemia, foi de extrema importância a indicação do subdomínio em que o termo se insere, realizada com a ajuda da criação da árvore de domínio, especificando assim o contexto em que a equivalência entre os idiomas se faz possível.

Em relação às três situações citadas por L'Homme (2020), que podem ocorrer ao considerar itens lexicais em duas línguas diferentes, conforme também mencionado no capítulo 2.1.3, foram observados os seguintes casos considerados relevantes, a serem descritos nos próximos parágrafos.

Em relação à primeira situação, de que um item lexical carrega no mínimo dois significados diferentes, estando o primeiro associado ao léxico de língua geral, e, o segundo, a um domínio especializado, foi observado o caso do termo *carta* (136 ocorrências). Enquanto, na língua geral, o significado mais comum é o de "correspondência, mensagem escrita ou impressa, que se envia a alguém, a uma instituição ou a uma empresa, para comunicar alguma coisa" (CARTA, 2021), no contexto do corpus de estudo, seu significado está relacionado apenas a uma "representação reduzida de determinada região, país, da superfície da Terra; mapa" (Ib. Idib.). Os trechos do corpus de estudo a seguir exemplificam essa utilização:

> Na guerra, porém, um fuzileiro naval (FN) em país estrangeiro pode não contar com a colaboração da população local e terá que se orientar com o único meio que em geral lhe estará disponível: a carta (BRASIL, 2020b, 16.1).

Ou, ainda:

> Uma carta é um desenho que não tem por finalidade reproduzir de forma fiel os acidentes naturais e artificiais da porção do terreno que representa, tal qual uma fotografia. Esses acidentes são representados por símbolos, de forma a facilitar o manuseio das cartas e padronizar sua confecção. (Ib. Idib., 16.2)

Já no corpus em inglês, não foi utilizada a palavra *chart*(0 ocorrências), mas a palavra *map*(36 ocorrências). Seu equivalente *chart*, derivado do vocábulo latino *charta*, com referência no grego, *chártes* (ETIMOLOGIA, 2021), é mais utilizado no inglês com o sentido de gráfico ou carta náutica (CHART, 2021; DIFFERENCE, 2021)[29]. Ou seja, ainda que possa ser considerado um tipo de mapa, um *chart* não

29. Do inglês, em CHART, 2021: "a drawing that shows information in a simple way, often using lines and curves to show amounts"; e, ainda, "a detailed map of an area of water: a naval chart".

é utilizado no inglês com o mesmo significado que em português, especialmente quando relacionado aos Fuzileiros Navais.

Tal utilização do termo *carta* no português pode estar associada à história e à formação do Fuzileiro Naval brasileiro, uma vez que esta se origina na Armada, ou seja, na Marinha, mas seria necessário um estudo terminológico diacrônico para analisar a evolução da difusão do uso do termo *carta* nessa área de especialidade.

Segunda situação: um item lexical transmite (pelo menos) dois significados diferentes ligados a diferentes áreas do conhecimento. Exemplo: *Cabo*, no contexto do CFN, é um termo que designa uma Praça mais antiga do que um Soldado, porém mais moderna do que um Sargento. Ainda, podemos falar em cabos de telecomunicações (do inglês, *wire*). Por isso, é importante a criação da árvore de domínio e a especificação do subdomínio a que o termo se encontra subordinado, eliminando assim a ambiguidade e a utilização errônea do equivalente tradutório.

Terceira situação: um item lexical carrega (pelo menos) dois significados diferentes que coexistem no mesmo domínio. Além do exemplo já mencionado em 2.1.3 em relação ao termo *comando* (no domínio militar: 1. ordem, instrução ou 2. conjunto de instâncias militares superiores), foi observado também o caso da palavra *segurança*, que possui equivalência parcial, ou seja, em inglês é feita uma distinção que não é feita em português (*safety* e *security*). *Safety* se refere à segurança física de pessoas individuais, como a proteção contra crimes, acidentes ou riscos à saúde[30]. Já *security* é um termo mais amplo, que pode estar relacionado à segurança de uma nação frente a inimigos externos ou de patrimônios, por exemplo[31]. Ou seja, esses dois significados coexistem no mesmo domínio apenas em inglês.

Outro caso semelhante de equivalência parcial se observa em relação aos termos *war* e *warfare*. Nos português, ambos se traduzem como *guerra*, mas em inglês é feita a seguinte distinção: *war* é um termo mais abrangente, que denomina a luta armada entre dois ou mais países[32]. Já *warfare* se refere ao ato de fazer uma guerra ou a itens associados a guerra, como armas, métodos, entre outros[33]. Este

30. Segundo SAFETY (2021): "a state in which or a place where you are safe and not in danger or at risk".
31. Segundo SECURITY (2021): "protection of a person, building, organization, or country against threats such as crime or attacks by foreign countries".
32. Do inglês: "a state in which or a place where you are safe and not in danger or at risk" (WAR, 2021).
33. Do inglês: "the activity of fighting a war, often including the weapons and methods that are used" (WARFARE, 2021).

também é um caso de polissemia na área militar que só ocorre em inglês, e não em português.

Em relação ao problema de equivalência entre postos e graduações do CFN e do USMC/Marinha americana, conforme mencionado na Introdução deste trabalho, a fim de elucidar ao usuário do léxico o melhor uso recomendado, optou-se por anexar uma tabela das principais patentes e sua equivalência entre as três forças. Além disso, foi feito um texto explicativo sobre como realizar a equivalência tradutória para o inglês dessas patentes, visto que essa é uma situação complexa que envolve diferenças importantes entre Oficiais e Praças do CFN, da US Navy e do USMC, devido às particularidades do CFN.

Conforme mencionado anteriormente, todos os militares, sejam eles Oficiais ou Praças, são admitidos no Corpo Fuzileiros Navais mediante concurso público (MEDEIROS, 2013; CHAGAS, 2021). Quando os Oficiais adentram a Escola Naval, todos pertencem à MB. Apenas ao final do segundo ano de curso, esses militares escolhem se farão parte do Corpo da Armada, do Corpo de Intendentes ou do Corpo de Fuzileiros Navais. Por outro lado, as Praças escolhem o Corpo a que farão parte já ao prestar o concurso público, pois há provas distintas para admissão nos diferentes Corpos da MB. Ou seja, enquanto os Oficiais pertencem primeiro apenas à Marinha do Brasil, e depois, caso assim escolham, ao CFN, as Praças pertencem ao CFN desde que se tornam militares.

Provavelmente por esse motivo, que poderia ser mais bem explicado por estudos de terminologia diacrônica, a relação de equivalência das patentes em inglês, conforme observado no corpus paralelo utilizado no estudo piloto e também de acordo com especialistas da área, é feita da seguinte forma: para as patentes de Oficiais do CFN, é estabelecida uma relação de equivalência com as patentes utilizadas pela US Navy. Já para as patentes de Praças do CFN, é estabelecida a equivalência diretamente com as patentes de Praças do USMC. As posições dos termos nas tabelas de patentes constantes na página do léxico, além das mesmas cores utilizadas para reforçar a equivalência entre elas (azul para Oficiais do CFN e da US Navy e amarelo para Praças do CFN e do USMC) e das notas informativas no fim da página, também mostram essa relação.

Nessas mesmas tabelas, foram inseridos também os termos que faltavam e resolvidas as inconsistências em relação ao material terminológico bilíngue já exis-

tente anteriormente (MARINOTTO, 2011; U.S. NAVY, 2015), conforme mencionado na introdução deste trabalho (atualização na grafia de acordo com o Novo Acordo Ortográfico, correções nos equivalentes terminológicos e inserção de patentes faltantes). Ao lado dos postos ou graduações para os quais não foram encontrados equivalentes terminológicos diretos no português, como *Warrant Officer* e *Guarda Marinha*, foi inserido um traço no campo correspondente. Também foram inseridas as siglas e abreviaturas correspondentes ao lado de cada patente, além de notas explicativas. Convém ressaltar também que para o termo *Praças*, apesar das Praças do CFN não serem alistadas, mas concursadas, conforme já mencionado, propôs-se, ainda assim, a utilização do equivalente *Enlisted*, por ser mais comum no inglês americano e sugerir uma associação direta com o significado da unidade lexical no português. Tais informações podem ser vistas na figura a seguir.

Marinha e CFN do Brasil	US Navy	USMC
Oficiais*	*Comissioned Officers*	*Comissioned Officers*
Almirante (Alte)	*Fleet Admiral (FADM)*	*General (Gen)*
Almirante de Esquadra (Alte Esq)	*Admiral (ADM)*	*General (Gen)*
Vice-Almirante (V Alte)	*Vice-Admiral (VADM)*	*Lieutenant General (LtGen)*
Contra-Almirante (C Alte)	*Rear Admiral Upper Half (RADM)*	*Major General (MajGen)*
-	*Rear Admiral Lower Half (RDML)*	*Brigadier General (BGen)*
Capitão de Mar e Guerra (CMG)	*Captain (CAPT)*	*Colonel (Col)*
Capitão de Fragata (CF)	*Commander (CDR)*	*Lieutenant Colonel (LtCol)*
Capitão de Corveta (CC)	*Lieutenant Commander (LCDR)*	*Major (Maj)*
Capitão-Tenente (CT)	*Lieutenant (LT)*	*Captain (Capt)*
Primeiro-Tenente (1T)	*Lieutenant, Junior Grade (LTJG)*	*First Lieutenant (1stLt)*
Segundo-Tenente (2T)	*Ensign (ENS)*	*Second Lieutenant (2ndLt)*
Guarda-Marinha (GM)	-	-

Marinha e CFN do Brasil	US Navy e USMC
-	*Warrant Officers***
-	*Warrant Officer (WO)*
-	*Chief Warrant Officer 2 (CWO2)*
-	*Chief Warrant Officer 3 (CWO3)*
-	*Chief Warrant Officer 4 (CWO4)*
-	*Chief Warrant Officer 5 (CWO5)*

Marinha e CFN do Brasil	US Navy	USMC
Praças***	*Enlisted*	*Enlisted*
Suboficial-Mor do CFN	*Master Chief Petty Officer of the Navy (MCPON)*	*Sergeant Major of the Marine Corps (SMMC)*
Suboficial-Mor (SOMor)	*Command Master Chief Petty Officer (CMC)*	*Sergeant Major (SgtMaj)*
Suboficial (SO)	*Master Chief Petty Officer (MCPO)*	*Master Gunnery Sergeant (technical)(MGSgt)*
Primeiro-Sargento (1SG)	*Senior Chief Petty Officer (technical) (SCPO)*	*Master Sergeant (technical) (MSgt)/ First Sergeant (administrative) (1stSgt)*
-	*Chief Petty Officer (CPO)*	*Gunnery Sergeant (GySgt)*
Segundo-Sargento (2SG)	*Petty Officer First Class (PO1)*	*Staff Sergeant (SSgt)*
Terceiro-Sargento (3SG)	*Petty Officer Second Class (PO2)*	*Sergeant (Sgt)*
Cabo (CB)	*Petty Officer Third Class (PO3)*	*Corporal (Cpl)*
-	*Seaman (SN)*	*Lance Corporal (LCpl)*
Soldado (SD) / Marinheiro (MN)	*Seaman Apprentice (SA)*	*Private First Class (PFC)*
Recruta	*Seaman Recruit (SR)*	*Private (PVT)*

Figura 26 – Extrato das tabelas de patentes militares bilíngue

Já os equivalentes terminológicos sugeridos para as Organizações Militares (OM) do CFN foram inseridos no organograma, conforme mostrado na figura a seguir.

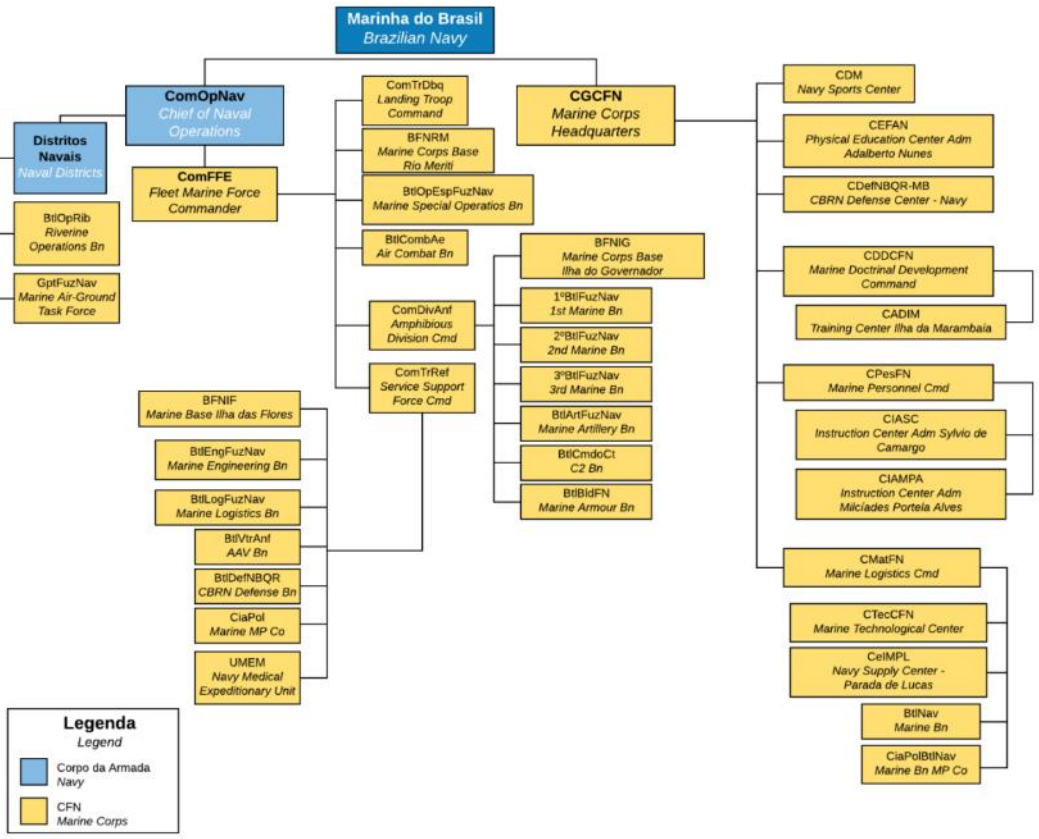


Figura 27 – Organograma bilíngue de Organizações Militares do CFN

Para alguns termos, não foi possível, em um primeiro momento, encontrar equivalentes terminológicos no corpus por meio do cruzamento de dados, como *flancoguarda* e *testa da vanguarda*. Nesses casos, foram consultados materiais terminológicos militares do Exército e do Departamento de Defesa norte-americanos (U.S. ARMY, 2015; DEPARTMENT OF DEFENSE, 2007) ou, ainda, a própria página oficial do USMC na internet (etapa 5.5). Em seguida, esse termo equivalente encontrado no material especializado era buscado no corpus para confirmação. Itens para os quais não foram encontrados equivalentes terminológicos no material consultado foram cortados do léxico. Itens que não foram encontrados no corpus de estudo em português, como alguns postos de Oficiais, mas que complementavam o respectivo organograma, foram buscados nos materiais terminológicos militares mencionados e acrescentados no léxico. Para os nomes de OM, como *Centro de Adestramento da Ilha da Marambaia*, que não possuem equivalência em inglês, foram sugeridas traduções possíveis.

Outro caso de não equivalência, conforme mencionado no capítulo 2.1.3 deste estudo, ocorrido quando não há, na língua meta, um equivalente adequado

para expressar o significado de um termo na língua fonte, ocorreu com o termo *adsumus*. Ainda que haja tradução para ele no português, a utilização do termo pelo CFN é mantida em latim, uma vez que ele é considerado, dessa forma, como o lema dos Fuzileiros Navais. Segundo BRASIL [s.d.], "ADSUMUS é o lema dos Fuzileiros Navais. É um termo de origem latina que significa 'Aqui estamos!', 'estar presente', 'estar junto' e, por extensão, significa um sentimento de permanente prontidão". A solução encontrada no léxico, por isso, foi de realizar o empréstimo direto do termo em latim como equivalente no inglês, adicionando essa informação como observação.

5.5
IDENTIFICAÇÃO DOS TERMOS NA ÁRVORE DE DOMÍNIO

Em seguida, cada termo foi identificado quanto às subdivisões da árvore de domínio. Para isso, foi incluída uma coluna na tabela de termos no Excel, ao lado da coluna *Variações*, uma vez que a grande quantidade de termos não permitiria inserir de forma organizada todos os termos na forma de organograma em uma árvore com todos os termos. Nessa coluna, foram identificados os respectivos domínios e subdomínios, conforme constam na árvore de domínio, ao lado de cada termo.

5.6
ORGANIZAÇÃO E REVISÃO DOS DADOS

Finalizada a organização dos termos quanto ao subdomínio dentro da árvore do CFN, foi a vez de realizar a organização dos dados quanto à macro e microestrutura do léxico. Além disso, foi realizada a revisão desses dados e algumas correções quando necessário. Nessa etapa, foi necessário também redigir pequenos textos explicativos sobre o léxico, a fim de fornecer informações importantes a respeito do material aos seus futuros usuários. Essas ações visaram a preparar o léxico para a posterior publicação na página da internet criada para esse fim.

Primeiramente, os dados foram organizados em relação à macroestrutura do léxico segundo os preceitos de Barros (2004) e Frübel (2006), ou seja, por ordem

alfabética. Também estão indicados ao lado dos termos suas áreas temáticas e subtemáticas, possibilitando assim a busca dos termos por categorias, conforme preconizado por Krieger e Finatto (2004), mais bem explicado no capítulo 2.1.5.

Assim, foram excluídas as colunas de avaliação dos termos quanto à Semântica Lexical da tabela do Excel. Os termos foram então organizados alfabeticamente (alguns termos, como OM e patentes, ainda estavam agrupados por campo semântico), diferenciando também letras maiúsculas e minúsculas. Uma vez que o código do site não aceita vírgulas ou pontos e vírgulas, foram utilizadas barras para a separação dos itens lexicais quando necessário. Barros (Ib. Idib.) e Frübel (2006) indicam que a ordem tradicional de organização dos verbetes é a alfabética. Krieger e Finatto (2004) apontam ainda que, além da ordem alfabética, a organização dos verbetes pode ser feita por área temática e subtemática, de acordo com a estrutura conceitual de um domínio, conforme análise de um especialista da área do conhecimento abrangida no material.

Quanto à microestrutura, ainda segundo Barros (Ib. Idib.), foi feita uma revisão final dos termos e do conjunto de informações referentes escolhidas para compor os verbetes do léxico. A tabela então manteve os seguintes campos: *Termos Pt-Br*, com os termos ordenados em português; *Variação*, com as variações lexicais dos termos em português, como siglas e abreviaturas, quando existentes; *Subárea*, identificando as subáreas conceituais dos termos, identificados pelas divisões da árvore de domínio do CFN; *Termos En*, com os equivalentes terminológicos em inglês; e *Variação*, com as variações lexicais em inglês. Os anexos (árvore de domínio, organograma de OM e tabela de patentes) foram revisados e salvos no formato .png para posterior carregamento no site.

5.7
PUBLICAÇÃO DO LÉXICO E ANEXOS EM PÁGINA NA INTERNET

A tabela e os anexos então foram publicados em uma página na internet, disponível no endereço https://maritraduz.com/lexico/. O léxico *Espírito de Corpus* se localiza dentro do domínio maritraduz.com, que a autora já possuía anteriormente. A página foi desenvolvida em linguagem PHP e MySql, usando o componente *jQuery DataTable* para facilitar as buscas ao digitar qualquer parte de qualquer termo constante na tabela.

Além da página em que se encontra o léxico, foram criadas três seções: Definição, Organizações Militares e Patentes. A seguir, serão mostradas imagens explicando cada uma delas. Por meio da figura 28, podemos ter uma visão geral da página, aberta em *Definição*.

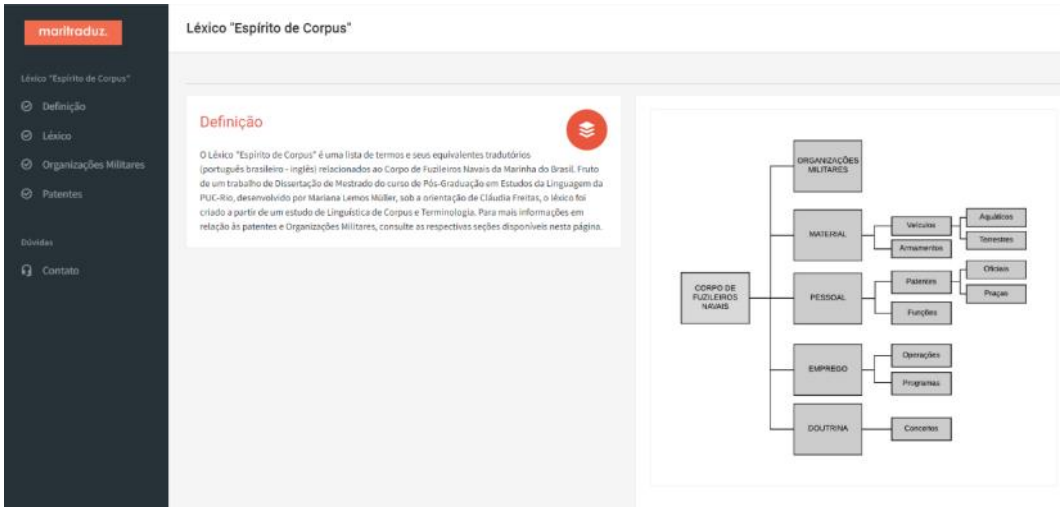


Figura 28 – Visão geral da página do léxico

Na seção *Definição*, foi publicado um pequeno texto introdutório que explica o que é o material. Ao lado desse texto, foi incluída a árvore de domínio dos termos cobertos pelo léxico.

Na página do léxico, além de rolar a página para cima e para baixo para possibilitar a busca dos termos desejados, é possível realizar uma busca rápida em uma caixa com essa finalidade. A imagem a seguir mostra a aparência da página com o campo de busca em destaque.

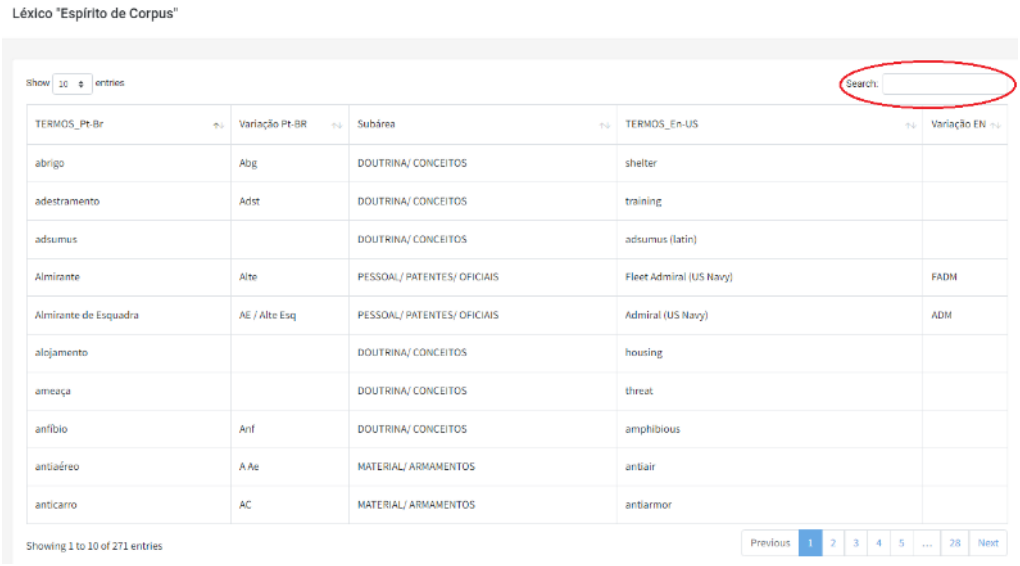


Figura 29 – Página do léxico com a caixa de busca em destaque

Ao realizar a busca pelo termo *Cabo* pela caixa de busca, por exemplo, a seguinte tela aparece ao usuário, mostrada pela figura 30.

Figura 30 – Resultado da busca pelo termo *Cabo*

Caso o usuário digite o nome de uma subárea da árvore de domínio no campo de busca, como *pessoal*, por exemplo, todos os termos pertencentes a essa subárea também aparecem ordenados alfabeticamente. A figura 31 exemplifica essa situação.

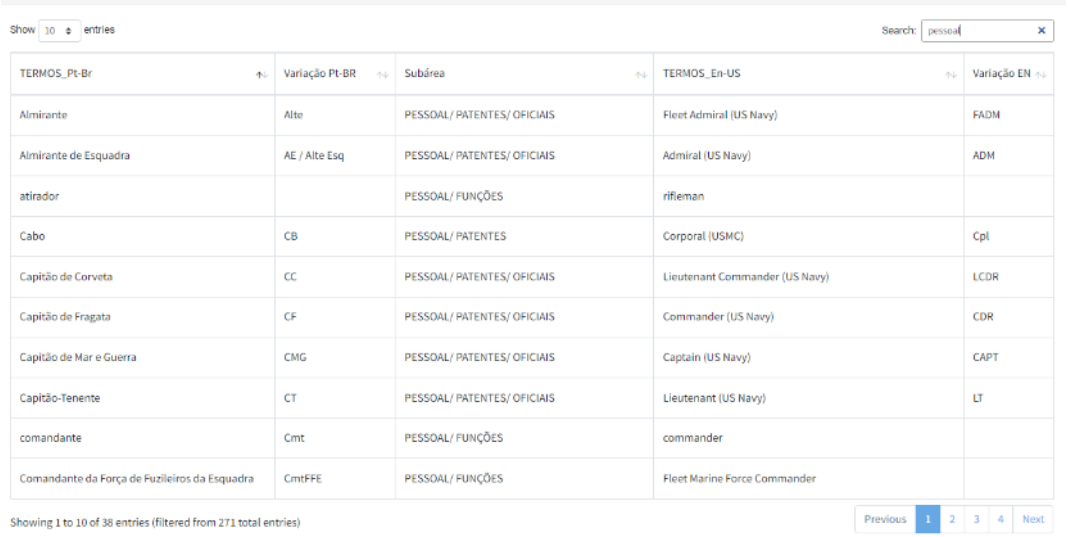


Figura 31 – Resultado da busca pela subárea *Pessoal*

Uma vez que a disposição por ordem alfabética dos termos, conforme mostrada acima, não possibilita estabelecer as relações hierárquicas entre as patentes, o usuário pode acessar a seção de *Patentes* para conseguir visualizar esses dados de forma mais estruturada por meio da tabela de patentes, mostrada com mais detalhes mais à frente.

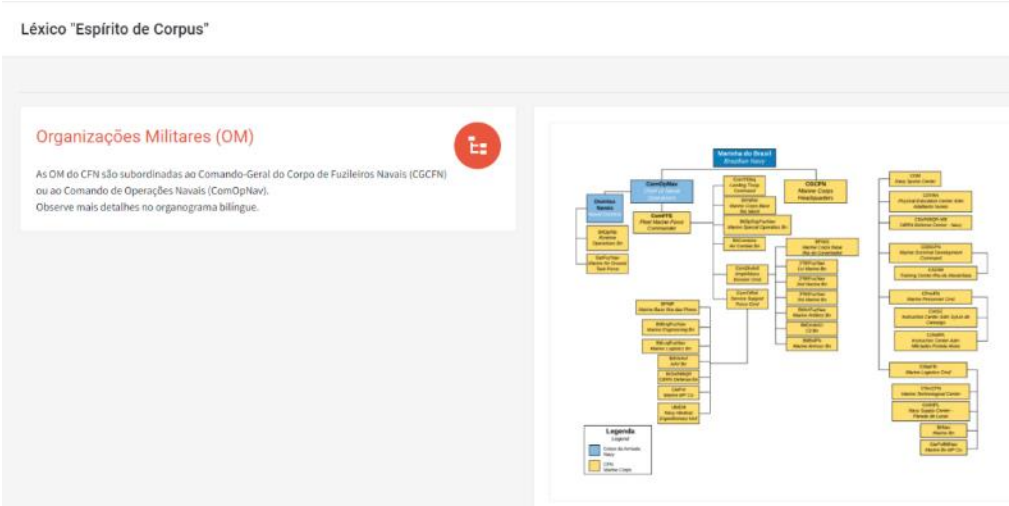


Figura 32 – Página de *Organizações Militares*

Em *Organizações Militares*, foi adicionado um breve texto explicativo sobre a subordinação das OM do CFN. Ao lado do texto, encontra-se o organograma bilíngue de OM, que pode ser ampliado para melhor visualização.

Já em *Patentes*, conforme ilustrado na imagem a seguir, foi disponibilizada a tabela das patentes do CFN organizadas hierarquicamente com os respectivos equivalentes terminológicos da US Navy e do USMC, ao lado de um breve texto explicativo. Esse texto orienta a utilização das patentes da US Navy como equivalentes terminológicos para as patentes de Oficiais da Marinha do Brasil e do CFN, conforme mencionado no capítulo 4.4. Já para as patentes de Praças, o texto sugere utilizar como referência os equivalentes do USMC.

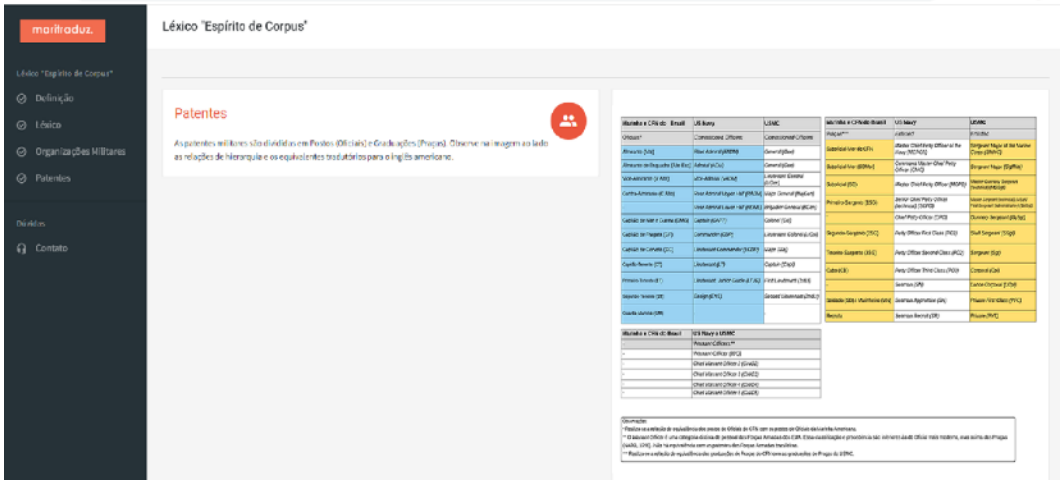


Figura 33 – Página de Patentes

Assim, foi concluída a publicação do léxico, que já está disponível para consulta online. Convém ressaltar que os termos constantes nesse léxico foram validados pelo Capitão de Fragata (FN) Maia, oficial de intercâmbio do CFN no USMC. Espera-se que esse material venha a auxiliar tanto na tradução de materiais relacionados ao CFN para o inglês americano, quanto na tradução de textos do USMC para o português brasileiro.

6
Considerações finais

Este estudo se propôs a criar um léxico bilíngue (português e inglês) de terminologia do CFN, área temática subordinada à área militar. Para isso, foram investigadas duas ferramentas de EAT: o AntConc 3.5.7 e o TermoStat Web 3.0, a fim de encontrar o melhor método. Por fim, foi aplicado um método híbrido (automático e manual) de extração e análise de terminologia bilíngue, por meio da extração de candidatos a termos de um corpus comparável da área de estudo com o auxílio da ferramenta AntConc, somada à aplicação da análise humana para validação. Como produto deste estudo, foi criado um léxico português-inglês de termos do CFN, o *Espírito de Corpus*.

A pesquisa se justifica uma vez que, apesar de haver diversas dúvidas e inconsistências em relação à tradução de termos do CFN do português para o inglês, devido às peculiaridades da força brasileira, não há no mercado hoje um material terminológico bilíngue voltado a essas particularidades. Outra justificativa se refere à necessidade de aprimorar o trabalho em relação à EAT que, apesar de trazer grande contribuição para a extração de material terminológico a partir da análise automática de corpora, apresenta ainda alguns problemas, como os relacionados a ruído e a silêncio. Outra dificuldade existente hoje que se buscou solucionar se refere à identificação dos itens lexicais que são realmente termos dentro das listas extraídas automaticamente e daqueles que são palavras de língua geral sem valor terminológico dentro de determinada área do conhecimento. Convém ressaltar, dentre suas limitações, que este trabalho não pretendeu resolver questões relacionadas à equivalência parcial.

A pesquisa se dividiu em duas fases principais: estudo piloto e confecção do *Espírito de Corpus*. O estudo piloto consistiu em uma série de testes realizados a fim de comparar o desempenho das duas ferramentas de Extração Automática de Termos. Primeiramente, os termos foram extraídos manualmente dos dois subcorpora em português e inglês, criando assim uma lista de referência, ou LR. A LR foi organizada alfabeticamente em duas colunas, compostas por termos em português e seus equivalentes terminológicos em inglês. Depois, foram feitos os testes com os dois subcorpora, iniciando com o TermoStat, por já apresentar as configurações

prontas, não permitindo ajustes. Em seguida, os testes foram realizados com o AntConc, cujas configurações foram ajustadas de forma que fossem aproximadas ao máximo das configurações do TermoStat (aproximação da linha de corte e utilização de corpus de referência de mesma tipologia).

Para a confecção do *Espírito de Corpus*, foi aplicado o método híbrido com a utilização da ferramenta que se destacou nos testes (AntConc). Assim, foram aproveitados os termos detectados no estudo piloto e buscados novos termos por meio da utilização de um corpus comparável bilíngue do CFN.

Os testes revelaram que o TermoStat apresentou os melhores índices para uma análise totalmente automática. Por isso, sugeriu-se seu uso para estudos que não permitam a realização de uma análise tão minuciosa, devido a limitações de tempo ou de pessoal, por exemplo.

Já os melhores resultados em geral foram obtidos ao realizar uma análise híbrida (automática e humana) com o AntConc sem qualquer linha de corte, realizando primeiramente a extração automática das palavras-chave e, em seguida, analisando-as em contexto, porque assim foram obtidos os maiores índices de precisão, abrangência e medida F. O teste sem a linha de corte não pôde ser realizado no TermoStat devido ao programa não possibilitar um ajuste manual das configurações, permitindo apenas que a análise humana corrija o ruído (eliminação de falsos positivos) da lista extraída. Logo, impossibilita a correção do silêncio (resgate de falsos negativos da lista de eliminados pela linha de corte) da EAT. Já o AntConc, por possibilitar o ajuste das configurações, permitiu que a análise humana corrigisse tanto o ruído quanto o silêncio (ao eliminar a linha de corte, pôde-se resgatar os falsos negativos), obtendo 100% de abrangência.

Em relação ao desempenho das ferramentas quanto aos idiomas dos subcorpora, foi observada uma lista de termos muito maior na EAT em português do que em inglês (3,76 vezes maior no AntConc e 5,19 vezes no TermoStat). Tal fato pôde ser explicado ao procurar termos militares de alta ocorrência nessa área de especialidade no corpus de referência em português, em que foram encontradas poucas ocorrências. Quando os equivalentes terminológicos desses mesmos termos de alta frequência foram procurados no corpus de referência em inglês, foi encontrada uma ocorrência 6.000% maior. Assim, constatou-se que a diferença de quantidade de termos extraídos nas duas línguas ocorreu pois os termos militares são mais frequentes nos textos jornalísticos estadunidenses; logo, o corpus de referência não

atribuiu a esses itens lexicais uma alta chavicidade a ponto de considerá-los termos. Dessa forma, a EAT em inglês apresentou uma lista muito menor de termos do que a lista em português. Foi constatada, assim, a importância da escolha criteriosa do corpus de referência para a melhora do desempenho da EAT.

Este trabalho forneceu as seguintes contribuições:

- sugestões de soluções aos problemas tradutórios identificados relativos ao CFN, como os relacionados a Postos e Graduações, aos nomes de Organizações Militares e ao uso de siglas e abreviaturas;
- a avaliação das ferramentas AntConc e TermoStat, o estudo das suas funcionalidades e o ajuste das suas configurações para aprimorar os resultados em relação à EAT e solucionar alguns problemas atribuídos em relação ao ruído e ao silêncio;
- a comparação do efeito da EAT nos idiomas inglês e português tanto nos corpora de estudo quanto nos corpora de referência;
- as descobertas em relação à adequação dos corpora de referência de origem jornalística;
- a compilação de um corpus paralelo bilíngue da subárea do CFN contendo cerca de 2.500 palavras e de um corpus comparável bilíngue composto por textos da mesma área de especialidade, com cerca de 180.000 palavras;
- o léxico *Espírito de Corpus*, que já está disponível na internet para consulta, com 270 termos em português e seus equivalentes terminológicos em inglês americano, contendo dados sobre a área de conhecimento do termo e variantes, além de informações adicionais como notas, abreviaturas e organogramas sobre hierarquia das patentes e organizações militares;
- o auxílio no procedimento de validação manual de termos extraídos automaticamente ao aplicar os critérios da Semântica Lexical segundo L' Homme (2020), que permitem verificar o caráter terminológico de determinado item lexical.

Por fim, esta pesquisa permitiu também demonstrar que, apesar da eficácia e rapidez das ferramentas de EAT, é imprescindível a avaliação humana das linhas de concordância a fim de se atingir um resultado excelente, daí a importância da metodologia híbrida aplicada.

Para futuros trabalhos, propõe-se ampliar o léxico por meio da detecção e inclusão de mais termos; apresentar definições, a fim de fornecer mais informações ao usuário do material, além de exemplos de uso extraídos de corpora, para propor sugestões que resolvam possíveis casos de equivalência parcial dos termos, por exemplo.

7
REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, G. M. B.; ALUÍSIO, S. M.; OLIVEIRA, L. H. M. O método em Terminologia: revendo alguns procedimentos. In: ISQUERDO, Aparecida Negri; ALVES, Ieda Maria. (Orgs.). **Ciências do léxico: lexicologia, lexicografia, terminologia.** 1 ed. Campo Grande/São Paulo: Editora da UFMS/Humanitas, 2007, v. III, p. 409-420.

ANDRADE, M. M. Lexicologia, terminologia: definições, finalidades, conceitos operacionais. In: OLIVEIRA, Ana Maria Pires de; ISQUIERDO, Aparecida Negri. **As ciências do léxico: Lexicologia, Lexicografia, Terminologia**. Campo Grande (MS): Ed. UFMS, 2001. 2ª ed.

ASTON, G. e KÜBLER, N. (2010) Corpora in translators training, in M. McCarthy and A. O'Keefe (eds.). **Routledge Handbook of Corpus Linguistics**. Routledge: London, 2010.

ANTHONY, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Disponível em <https://www.laurenceanthony.net/software>

BAKER, M. **Corpora in translation studies. An overview and suggestions for future research.** Target 7(2), 1995, p. 223-243.

_____. A corpus-based view of similarity and difference in translation. In: ARDUINI, S.; HODGSON, R. (Eds.). **Translating Similarity and Difference**. Manchester: St. Jerome, 2004, p. 1-18.

BAKER, P. **Sociolinguistics and corpus linguistics**. Edinburgh: Edinburgh. University Press, 2010.

BARROS, L. A. **Curso básico de terminologia**. São Paulo: Edusp, 2004.

BIBER, D. et al. **Corpus linguistics Investigating language structure and use**. Cambridge: Cambridge University Press, 1998.

BOUTIN-QUESNEL, R. et al. **Vocabulaire systématique de la terminologie**. Québec: Publications du Québec, 1985.

BRASIL. Ministério da Defesa. **Glossário das Forças Armadas**. MD35-G-01. 2015. 5a edição.

_____. Diretoria de Comunicações e Tecnologia da Informação da Marinha. **Boletim de Ordens e Notícias nº 120 de 17 de fevereiro de 2016**. Disponível em: https://www.marinha.mil.br/sites/www.marinha.mil.br.dadm/files/bono120E.pdf
Acesso em: 18 de agosto de 2021.

_____. Ministério da Defesa. **Política de Defesa e Estratégia Nacional de Defesa**. 2020a.

_____. Comando-Geral do Corpo de Fuzileiros Navais. CGCFN-3101.1 (Ed. 2). **Manual Básico do Fuzileiro Naval**. 2020b.

_____. Ministério da Defesa. **Organograma das Forças Armadas Brasileiras.** 2021a. Disponível em: <https://www.gov.br/defesa/pt-br/acesso-a-informacao/institucional-2/estrutura-organizacional/organograma.pdf>. Acesso em: 15 de julho de 2021.

_____. **Postos e Graduações**. 2021b. Disponível em: <https://www.marinha.mil.br/postos-e-graduacoes>. Acesso em: 18 de agosto de 2021.

_____. Marinha do Brasil. **Corpo de Praças Fuzileiro Naval**. 2021c. Disponível em: <https://www.marinha.mil.br/ensino/?q=carreira-naval/pracas-corpo-fn>. Acesso em: 18 de agosto de 2021.

_____. Ministério da Defesa. **Curriculum vitae do Vice-Almirante Carlos Chagas**. 2021d. Disponível em: <https://www.gov.br/defesa/ptbr/arquivos/lai/institucional/curriculo/gm/Curriculo_VA_Carlos_Chagas_Vianna_Braga_atual.pdf>. Acesso em: 11 de novembro de 2021.

_____. Ministério da Defesa. **Manual de Abreviaturas, Siglas, Símbolos e Convenções Cartográficas das Forças Armadas**. MD33-M-02, 2008. Disponível em: <https://www.gov.br/defesa/ptbr/arquivos/File/legislacao/emcfa/publicacoes/md33a_ma_02a_mnla_abreva_siglaa_sblcnvca_crtgrffaa_3aed2008.pdf> . Acesso em: 28 de dezembro de 2021.

_____. **Valores**. [s.d.] Disponível em: https://www.marinha.mil.br/cgcfn/valores. Acesso em: 28 de dezembro de 2021.

BUSSMANN, Hadumod. **Routledge dictionary of language and linguistics**. London & New York: Routledge, 2006. (formato e-book)

_____. La neologia efímera. **Miscellània Joan Bastardas.** Barcelona: Publicacions de l'Abadia de Montserrat, 1989, p. 37-58.

_____. **La terminología. Teoria, metodologia, aplicaciones**. Barcelona: Editorial Antártida/Empúries, 1993.

CABRÉ, M. T. **Terminology: Theory, methods and applications**. Amsterdam/Philadelphia: John Benjamins Publishing, 1999.

CANÇADO, M. **Manual de Semântica. Noções básicas e exercícios.** São Paulo: Contexto, 2013.

CATFORD, J. C. **A Linguistic Theory of Translation: an Essay on Applied Linguistics.** London: Oxford University Press, 1965.

CHAGAS, C. **BRAZILIAN MARINE CORPS: National Strategic Amphibious Expeditionary Force in Readiness**. 19 de maio de 2021. Apresentação de PowerPoint. 134 slides, color.

CARTA. In: DICIO, Dicionário online de Português, 2021. Disponível em: https://www.dicio.com.br/carta/. Acesso em: 20 de dezembro de 2021.

CETENFolha. In: LINGUATECA, 2018. Disponível em: https://www.english-corpora.org/time/. Acesso em: 31 de julho de 2021.

CHART. In: CAMBRIDGE Dictionary, 2021. Disponível em: https://dictionary.cambridge.org/pt/dicionario/ingles/chart. Acesso em: 20 de dezembro de 2021.

CHUNG, T. M. A corpus comparison approach for terminology extraction. In: **Terminology**, Volume 9, Number 2, p. 221-246. John Benjamins Publishing Company, 2003.

CONRADO, M. S., PARDO, T. A. S. e REZENDE, S. O. **A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set***. Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMCUSP), 2020. Disponível em: <https://sites.icmc.usp.br/taspardo/NAACL-HLT2013-ConradoEtAl.pdf>. Acesso em: 07 de agosto de 2020.

DAYRELL, C. **O uso de corpora para o estudo da tradução: objetivos e pressupostos**. Tradução em Revista, Departamento de Letras/Puc-Rio, v. 2, p. 87-102, 2005.

DIFFERENCE between maps and charts. In: differencebetween.net. Disponível em: http://www.differencebetween.net/miscellaneous/geography-miscellaneous/difference-between-maps-and-charts/. Acesso em: 20 de dezembro de 2021.

DROUIN, P. Term extraction using non-technical corpora as a point of leverage. In: **Terminology**, 2003. Disponível em: <https://www.researchgate.net/publication/228683045_Term_extraction_using_nontechnical_corpora_as_a_point_of_leverage.> Acesso em: 07 de agosto de 2020.

_____. **User's Guide TermoStat 3.0**, 2010. Disponível em: <http://termostat.ling.umontreal.ca/doc_termostat/doc_termostat.html>. Acesso em: 07 de agosto de 2020.

ESTOPÀ BAGOT, R. "Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)", 1999, Tese de Doutorado. Universidade Pompeu Fabra.

ESTOPÀ BAGOT, R. "Extracción de Terminologia: elementos para la construcción de un extractor", TradTerm 7, 2001, Revista do Centro Interdepartamental de Tradução e Terminologia FFLCH - USP, p. 225-50.

ETIMOLOGIA de carta. In: **ETIMOLOGIA Origem do conceito**. Disponível em: < https://etimologia.com.br/carta/>. Acesso em: 28 de dezembro de 2021.

FILLMORE C. J. **Frames semantics and the nature of language. In: Conference on the origin and development of language and speech**. New York: New York Academy of Science.Vol. 280, p. 20-32, 1976. Disponível em: <http://dx.doi.org/10.1111/j.1749-6632.1976.tb25467.x> Acesso em: 10 de maio de 2022.

FRÜBEL, A.C.M. **Glossário de neologismos terminológicos da saúde humana: uma contribuição para a descrição do léxico corrente do português do Brasil**. 2006. Tese (Doutorado em Linguística e Língua Portuguesa) - Universidade Estadual Paulista Júlio de Mesquita Filho, São Paulo.

ISO - INTERNATIONAL ORGANISATION FOR STANDARDISATION. **ISO/DIS 1087 - 1.2 Terminology work - Vocabulary**. Geneve: ISO, 2000.

_____. **ISO/DIS 5127 - Information and documentation — Foundation and vocabulary**. Geneve: ISO, 2017.

KENNY, D. **Lexis and Creativity in Translation: A Corpus-based Study.** London and New York: Routledge, 2001

KENNING, M. 2010. 'What are parallel and comparable corpora and how can we use them?' In: A. O'Keffee and M. McCarthy (eds.), **The Routledge handbook of corpus linguistics**, 487–501.

KRIEGER, M. G. "Sobre Terminologia e seus objetos". In: KRIEGER, M. G.; MACIEL, A. M. B. (orgs) **Temas de Terminologia**. São Paulo: Humanitas/FFLCH/USP, 2001.

_____, M. G.; FINATTO, M. J. B. **Introdução à Terminologia: teoria e prática**. São Paulo: Contexto, 2004.

LABATE, F. G. **Vocabulário da Economia: Formas de apresentação dos estrangeirismos**. São Paulo: FFLCH- USP, 2008. Dissertação (Mestrado em Filologia e Língua Portuguesa).

LAFACE, A. "Definição do vocabulário terminológico no universo acadêmico: reflexões didático-pedagógicas" In: OLIVEIRA, Ana Maria Pires de; ISQUIERDO, Aparecida Negri. **As ciências do léxico: Lexicologia, Lexicografia, Terminologia**. Campo Grande (MS): Ed. UFMS, 2001. 2ª Ed.

L'HOMME, M. C. **La terminologie : principes et techniques**, Montréal, Les Presses de **l**'Université de Montréal, Coll. « Paramètres », 278 p., 2004.

_____. **Lexical Semantics for Terminology: An introduction.** Amsterdam / Philadelphia: John Benjamins Publishing Company, 2020.

LOPES, L., FERNANDES, P., e VIEIRA, R. **Domain term relevance through tf-dcf. In Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012)**, pages 1001–1007. Las Vegas, USA: CSREA Press, 2012.

LOPES, L. e VIEIRA, R. Aplicando Pontos de Corte para Listas de Termos Extraídos. In: Proceedings of STIL 2013, 2013.

LOPES, L., VIEIRA, R., FINATTO, M., MARTINS, D., ZANETTE, A., & JUNIOR, L. (2009). Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde, 3**(1). doi:https://doi.org/10.3395/reciis.v3i1.821

MACIEL, A. M. B. Terminologia, linguagem de especialidade e dicionários in KRIEGER, Maria da Graça. MACIEL, Anna Maria Becker. (Org.) **Temas de Terminologia**. Porto Alegre/São Paulo: Ed. Universidade/UFRGS e Humanitas/USP, 2001.

MARINOTTO, D. Dicionário de Expressões e Termos Militares - Inglês / English - Português / Portuguese. Curitiba: Editora CRV, 2011. 1ª Ed.

McENERY, A., XIAO, R. e TONO, Y. **Corpus-Based Language Studies: An Advanced Resource Book**. Nova York: Routledge, 2006.

McENERY, T. e HARDY, A. **Corpus Linguistics: Method, Theory and Practice**. Cambridge: Cambridge University Press, 2012.

MEDEIROS, A. C. e OLIVEIRA, L. C. **Fuzileiros Navais: confie neles!** Rio de Janeiro: Euangellus Comunicação, 2013.

MEIOS. In: Dicionário PRIBERAM, 2021. Disponível em: https://dicionario.priberam.org/meio. Acesso em: 18 de agosto de 2021.

MILITARY.COM. **Military Ranks: Everything You Need to Know**, 2021. Disponível em: rankshttps://www.military.com/join-military/military-ranks-everything-you-need-know.html. Acesso em: 18 de agosto de 2021a.

MILITARY.COM. **Enlisted Marine Corps Ranks**, 2021. Disponível em:https://www.military.com/marine-corps/enlisted-ranks.html#nco-ranks. Acesso em: 18 de agosto de 2021b.

MÜLLER, M. L. Estudo terminológico bilíngue de termos militares da subárea do Corpo de Fuzileiros Navais. Monografia (Pós-graduação em Técnicas, Práticas e Estudos da Tradução em Inglês – Português) – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro, Rio de janeiro, p. 53, 2019.

NATO. **STANAG 2116 - NATO codes for grades of military personnel**, Edition 4, 1996. Disponível em: <https://militaria.lv/stanag.htm>. Acesso em: 18 de agosto de 2021.

NIDA, E. **Language, Culture, and Translating.** Shanghai: Shanghai Foreign Language Education Press, 1993.

O'KEEFFE, A. e MCCARTHY, M. **The Routledge Handbook of Corpus Linguistics**. London and New York: Routledge, 2012.

OXFORD Languages. **Militar**. Disponível em: https://www.google.com/search?q=militar+defini%C3%A7%C3%A3o&oq=militar&aqs=chrome.1.69i59l2j69i57j35i39j0i131i433i512l2j69i61j69i60.3693j1j4&sourceid=chrome&ie=UTF-8. Acesso em: 6 de agosto de 2021.

PAIVA, P. T. P.; CAMARGO, D. C.; XATARA, C. M. **Uma reflexão sobre a elaboração de um léxico bilíngüe preliminar na subárea de cardiologia a partir do uso de termos encontrados em um corpus paralelo e em dois corpora comparáveis.** DELTA. Documentação de Estudos em Lingüística Teórica e Aplicada, v. 24:1, p. 1-22, 2008.

PAVEL, S. e NOLET, D. **Manual de Terminologia**. Adaptação para língua portuguesa por Enilde Faulstich. Canadá: Ministério de Obras Públicas e Serviços Governamentais do Canadá, 2002.

PIMENTEL, J. Adding a new language version to a lexical resource. Is it possible to assign term equivalents semi-automatically? In: **Terminàlia** 11, p. 20-29, 2013.

PYM, A. **Explorando Teorias da Tradução.** São Paulo: Perspectiva, 2017.

REY, A. A terminologia entre a experiência da realidade e o comando dos signos. In: ISQUIERDO, Aparecida Negri; ALVES, Ieda Maria (orgs). **As Ciências do Léxico. Lexicologia, Lexicografia e Terminologia**. Volume III. Campo Grande / São Paulo: EdUFMS/Humanitas, 2007.

SAFETY. In: CAMBRIDGE Dictionary, 2021. Disponível em: https://dictionary.cambridge.org/pt/dicionario/ingles/safety. Acesso em: 20 de dezembro de 2021.

SARDINHA, T. B. **Linguística de corpus: histórico e problemática**. D.E.L.T.A., São Paulo, EDUC, Vol. 16, n. 2, 2000.

_____. **Tamanho de Corpus**. In: the ESP, São Paulo, vol. 23, nº 2, p. 103-122. LAEL, PUCSP, 2003. Disponível em: <https://revistas.pucsp.br/index.php/esp/article/view/9381>. Acesso em: 20 de julho de 2021.

SECURITY. In: CAMBRIDGE Dictionary, 2021. Disponível em: <https://dictionary.cambridge.org/pt/dicionario/ingles/security>. Acesso em: 20 de dezembro de 2021.

SINCLAIR, J. 1991. **Corpus, Concordance, Collocation**. Oxford, UK, Oxford University Press.

_____. 2005. "Corpus and Text - Basic Principles" in **Developing Linguistic Corpora: a Guide to Good Practice**, ed. M. Wynne. Oxford: Oxbow Books. Disponível em: <http://ahds.ac.uk/linguistic-corpora/> Acesso em: 20 de agosto de 2021.

STUBBS, M. **Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture**. Oxford: Blackwell, 1996.

TAGNIN, S. E. O. **Os Corpora: instrumentos de auto-ajuda para o tradutor**. Cadernos de Tradução, Florianópolis, v. 1, n. 9, 191-219, 2002/1.

_____. A produção de glossários direcionados pelo corpus e orientados ao tradutor como metodologia de formação de tradutores. In: X Encontro Nacional de Tradutores / IV Encontro Internacional de Tradutores, 2009, Ouro Preto, Minas Gerais, Bras. **Anais do X Encontro Nacional de Tradutores / IV Encontro Internacional de Tradutores**. Ouro Preto, MG, Brasil: Editora da UFOP, 2009. Disponível em: http://docplayer.com.br/62545180-Anais-do-x-encontro-nacional-de-tradutores-iv-encontro-internacional-de-tradutores-abrapt-ufop-ouro-preto-de-7-a-10-de-setembro-de-2009.html. Acesso em: 31 de julho de 2021.

TELINE, M. F. ; ALMEIDA, G. M. B. ; ALUíSIO, S. M. Extração manual e automática de terminologia: comparando abordagens e critérios . In: **TIL 2003 - Evento Integrante do 16th Brazilian Symposium on Computer Graphics and Image Processing -SIBGRAPI 2003**, 2003, São Carlos. Proceedings of the 16th Brazilian Symposium on Computer Graphics and Image Processing, 2003. v. 1. p. 1-12.

TEIXEIRA, R. B. S. **Termos de (Onco)mastologia: uma abordagem mediada por corpus**. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo. São Paulo, p. 363. 2010.

TERMOSTAT Web Tutorial, Level I, in LinguisTech, 03 de agosto de 2020. Disponível em: <https://linguistech.ca/TermoStat_E_TUTCERTT_I_PartIr>. Acesso em: 03 de agosto de 2020.

TIME Magazine Corpus of America. [s.d.]. Disponível em: https://www.english-corpora.org/time/. Acesso em: 31 de julho de 2021.

TOGNINI-BONELLI, E. Working with corpora across languages. In: TOGNINI-BONELLI, E. **Corpus Linguistics at work**. Amsterdam/Atlanta, John Benjamins, 2001.

TOURY, G. **Descriptive Translation Studies and Beyond**. Amsterdam: John Benjamin, 1995.

U.S. ARMY. **Combined Arms Center (CAC). Military Review - English-Portuguese Dictionary of Military Terminology (A-Z)**. Fort Leavenworth, Kansas, 2015.

WAR. In: CAMBRIDGE Dictionary, 2021. Disponível em: <https://dictionary.cambridge.org/pt/dicionario/ingles/war>. Acesso em: 20 de dezembro de 2021.

WARFARE. In: CAMBRIDGE Dictionary, 2021. Disponível em: <https://dictionary.cambridge.org/pt/dicionario/ingles/warfare>. Acesso em: 20 de dezembro de 2021.

8
Apêndices

TERMOS SIMPLES_PT			
1	ADSUMUS	26	FORÇA
2	AMEAÇA	27	GPTOPFUZNAV
3	ANFÍBIO	28	GRUPAMENTO
4	ATAQUE	29	GUERRA
5	BATALHÃO	30	MANOBRA
6	BTL	31	MANOBRAR
7	CASC	32	MATERIAL
8	CCA	33	MEIOS
9	CCT	34	MISSÃO
10	CFN	35	NAVAL
11	CIAB	36	OPERAÇÃO
12	COMANDANTE	37	OPERACIONAL
13	COMBATE	38	OPERAR
14	DEFESA	39	PELOTÃO
15	DESEMBARQUE	40	PESSOAL
16	DOUTRINA	41	PIRATARIA
17	EFETIVO	42	PREPARAR
18	EMA-302	43	PREPARO
19	EMBARCAÇÃO	44	PRONTIDÃO
20	EMBARQUE	45	RIBEIRINHA
21	EMPREGAR	46	SEGURANÇA
22	EMPREGO	47	SUBUNIDADE
23	ENC	48	TERRA
24	EXÉRCITO	49	TERRORISMO
25	EXPEDICIONÁRIO	50	TROPA

Apêndice 1 – Lista de referência de unigramas em português

TERMOS COMPOSTOS/COMPLEXOS_PT					
1	1 BTL INFANTARIA	33	CIA DE POLÍCIA	65	LABORATÓRIO NBQR
2	2+ BTL INFANTARIA	34	COMANDANTE DA MARINHA	66	MARINHA DO BRASIL
3	3 BTL INFANTARIA	35	COMANDANTE DE OPERAÇÕES NAVAIS	67	MISSÃO DE PAZ
4	AÇÃO CÍVICO-SOCIAL	36	COMANDO DA DIVISÃO ANFÍBIA	68	MÍSSEIS SUPERFÍCIE-AR
5	ÁGUAS JURISDICIONAIS BRASILEIRAS	37	COMANDO DE OPERAÇÕES NAVAIS	69	OPERAÇÃO "VERDE BRASIL" 1 E 2
6	AMAZÔNIA AZUL	38	COMANDO E CONTROLE	70	OPERAÇÃO AMAZÔNIA AZUL
7	ASSISTÊNCIA HUMANITÁRIA	39	COMPONENTE DE COMANDO	71	OPERAÇÃO ANFÍBIA
8	ATIVIDADE BENIGNA	40	COMPONENTE DE COMBATE AÉREO	72	OPERAÇÃO CONJUNTA
9	BASE DO RIO MERITI	41	CONJUGADO ANFÍBIO	73	OPERAÇÃO COVID-19
10	BATALHÃO DE ARTILHARIA	42	CORPO DE FUZILEIROS NAVAIS	74	OPERAÇÃO DE ASSISTÊNCIA HUMANITÁRIA
11	BATALHÃO DE BLINDADOS	43	DEFESA NBQR	75	OPERAÇÃO DE DESASTRE NATURAL
12	BATALHÃO DE COMBATE AÉREO	44	DESMINAGEM HUMANITÁRIA	76	OPERAÇÃO DE GUERRA NAVAL
13	BATALHÃO DE ENGENHARIA	45	DIRETORIA DE HIDROGRAFIA	77	OPERAÇÃO DE INFORMAÇÃO
14	BATALHÃO DE OPERAÇÕES ESPECIAIS	46	DISTRITO NAVAL	78	OPERAÇÃO DE PAZ
15	BATALHÃO DE VIATURAS ANFÍBIAS	47	DIVISÃO ANFÍBIA	79	OPERAÇÃO HUMANITÁRIA
16	BATALHÃO LOGÍSTICO	48	DOUTRINA MILITAR NAVAL	80	OPERAÇÃO INTERAGÊNCIA
17	BATALHÃO DE INFANTARIA	49	ELEMENTO ANFÍBIO	81	OPERAÇÕES ESPECIAIS
18	BATALHÃO DE OPERAÇÕES RIBEIRINHAS	50	EMBARCAÇÃO DE GUERRA	82	OPERAÇÕES RIBEIRINHAS
19	BRIGADA ANFÍBIA	51	EMPREGO LIMITADO DA FORÇA	83	PATRULHA NAVAL
20	BTL COMANDO E CONTROLE	52	EQUIPE DE SEGURANÇA	84	PODER NAVAL
21	BTL DE ARTILHARIA	53	ESTRATÉGIA NACIONAL DE DEFESA	85	PROGRAMA DE DESENVOLVIMENTO SOCIAL
22	BTL DE BLINDADOS	54	ESTUDO TOPOTÁTICO DO TERRENO	86	PROGRAMA FORÇAS NO ESPORTE
23	BTL DE COMBATE AÉREO	55	FORÇA AERONAVAL	87	PROJEÇÃO DE PODER
24	BTL DE DEFESA NBQR	56	FORÇA ANFÍBIA E EXPEDICIONÁRIA	88	PRONTO EMPREGO
25	BTL DE ENGENHARIA	57	FORÇA DE SUBMARINOS	89	SEGURANÇA MARÍTIMA
26	BTL DE INFANTARIA	58	FORÇA DISTRITAL	90	SISTEMA DE ARMAS
27	BTL DE LOGÍSTICA	59	FORÇA NAVAL	91	TREINAMENTO FÍSICO MILITAR
28	BTL DE OPERAÇÕES ESPECIAIS	60	FUZILEIRO NAVAL	92	TROPA DE DESEMBARQUE
29	BTL DE OPERAÇÕES RIBEIRINHAS	61	GRUPO DE ASSESSORAMENTO	93	TROPA DE EMBARQUE
30	BTL DE VIATURAS ANFÍBIAS	62	GUARDA COSTEIRA	94	TROPA DE REFORÇO
31	CENÁRIO ESTRATÉGICO DE INTERESSE	63	GUERRA CIBERNÉTICA	95	UNIDADE ANFÍBIA
32	CENTRO DE DEFESA NBQR	64	GUERRA HÍBRIDA	96	UNIDADE MÉDICA EXPEDICIONÁRIA
				97	VICE-ALMIRANTE (FN)

Apêndice 2 – Lista de referência de bigramas, trigramas e quadrigramas em português

TERMOS SIMPLES_EN			
1	AAV	27	MILITARY
2	ADSUMUS	28	MISSION
3	AMPHIBIOUS	29	NAVAL
4	ARMY	30	NEO
5	ATTACK	31	OPERATING
6	BATTALION	32	OPERATION
7	BN	33	OPERATIONAL
8	BOARDING	34	PERSONNEL
9	CBRN	35	PIRACY
10	COMBAT	36	PLATOON
11	COMMANDANT	37	PREPARE
12	COMMANDER	38	READINESS
13	DEFENSE	39	RIVERINE
14	DOCTRINE	40	RU
15	EMPLOY	41	SAFETY
16	EMPLOYMENT	42	SECURITY
17	EXPEDITIONARY	43	SPMAGTF
18	GCE	44	STRENGTH
19	LAND	45	TERRORISM
20	LANDING	46	THREAT
21	LCE	47	TRAINING
22	MAGTF	48	TROOP
23	MANEUVER	49	VESSEL
24	MARINE	50	WAR
25	MARITIME	51	WARFARE
26	MATERIEL		

Apêndice 3 – Lista de referência de unigramas em inglês

TERMOS COMPLEXOS_EN					
1	1 INFANTRY BN	36	FORCES IN SPORTS PROGRAM	71	NAVAL FORCE
2	1ST RIVERINE OPERATIONS BATTALION	37	GROUND COMBAT ELEMENT	72	NAVAL OPERATIONS COMMAND
3	2 + INFANTRY BN	38	HAVE MARINES PERMANENTLY READY	73	NAVAL PATROLLING
4	2ND RIVERINE OPERATIONS BATTALION	39	HUMANITARIAN ASSISTANCE	74	NAVAL POWER
5	3RD INFANTRY BATTALION	40	HUMANITARIAN ASSISTANCE OPERATION	75	NAVAL SPECIAL OPERATIONS COMMAND
6	3RD RIVERINE OPERATIONS BATTALION	41	HUMANITARIAN DEMINING	76	NAVAL WARFARE OPERATION
7	AIR COMBAT BATTALION	42	HUMANITARIAN OPERATION	77	NAVY MEDICAL EXPEDITIONARY UNIT
8	AIR COMBAT ELEMENT	43	HYBRID WARFARE	78	NAVY STRATEGIC PLAN 2040
9	AMPHIBIOUS AND EXPEDITIONARY FORCE	44	HYDROGRAPHIC OFFICE	79	NBCR DEFENSE
10	AMPHIBIOUS ASSAULT VEHICLE	45	INFANTRY BATTALION	80	NBCR LAB
11	AMPHIBIOUS DIVISION	46	INFORMATION OPERATION	81	NONCOMBATANT EVACUATION
12	AMPHIBIOUS DIVISION COMMAND	47	INTERAGENCY OPERATION	82	OPERATION BLUE AMAZON
13	AMPHIBIOUS OPERATION	48	JOINT OPERATION	83	OPERATION COVID-19
14	BASE RIO MERITI	49	LAND OPERATION	84	ORGANIZATION FOR COMBAT
15	BENIGN ACTIVITY	50	LANDING TROOP	85	PEACEKEEPING MISSION
16	BLUE AMAZON	51	LANDING TROOP COMMAND	86	PEACEKEEPING OPERATION
17	BOARDING TROOP	52	LAW AND ORDER ENFORCEMENT	87	PHYSICAL TRAINING
18	BRAZILIAN JURISDICTIONAL WATERS	53	LIMITED USE OF FORCE	88	PLACE OF STRATEGIC INTEREST
19	BRAZILIAN NAVY	54	LOGISTICS COMBAT ELEMENT	89	PORTUGUESE ROYAL MARINE BRIGADE
20	BRAZILIAN NAVY AMPHIBIOUS CONJUGATE	55	MARINE ADVISORY GROUP	90	POWER PROJECTION
21	CBRN DEFENSE	56	MARINE AMPHIBIOUS BRIGADE	91	RECOVERY AND HOSTAGE RESCUE
22	CHIEF OF NAVAL OPERATIONS	57	MARINE AMPHIBIOUS UNIT	92	REGION UNIT
23	CIVIC-SOCIAL ACTION	58	MARINE ARMOUR BATTALION	93	RIVERINE OPERATION
24	COAST GUARD	59	MARINE ARTILLERY BATTALION	94	RIVERINE OPERATIONS BATTALIONS
25	COMAND AND CONTROL BATTALION	60	MARINE CORPS	95	SEARCH AND SEIZURE
26	COMBATANT PERFORMANCE LAB	61	MARINE CORPS BASE GOVERNADOR	96	SECURITY TEAMS
27	COMMAND AND CONTROL	62	MARINE CORPS BASE RIO MERITI	97	SERVICE SUPPORT FORCE
28	COMMAND ELEMENT	63	MARINE CORPS HEADQUARTERS	98	SERVICE SUPPORT FORCE COMMAND
29	COMMANDER OF THE NAVY	64	MARINE CORPS STRATEGIC CONCEPT	99	SOCIAL DEVELOPMENT PROGRAM
30	CYBER WARFARE	65	MARINE ENGINEERING BATTALION	100	SPECIAL OPERATION
31	DISASTER RELIEF OPERATION	66	MARINE LOGISTICS BATTALION	101	SPECIAL OPERATIONS BATTALION
32	DISTRICT FORCE	67	MARINE MP CO	102	SPMAGTF
33	FLEET MARINE FORCE	68	MARITIME SAFETY	103	SUBMARINE FORCE
34	FLEET MARINE FORCE COMMAND	69	NAVAL AIR FORCE	104	TOPOTACTIC STUDY OF TERRAIN
35	FLEET MARINE FORCE COMMANDER	70	NAVAL DISTRICT	105	VICE ADMIRAL
				106	WAR VESSEL
				107	WEAPON SYSTEM

Apêndice 4 – Lista de referência de bigramas, trigramas e quadrigramas em inglês

UNIGRAMAS ANTCONC_PT		UNIGRAMAS TERMOSTAT_PT	
1	NAVAL	1	NAVAL
2	BTL	2	BTL
3	FORÇA	3	ANFÍBIO
4	ANFÍBIO	4	GRUPAMENTO
5	EXPEDICIONÁRIO	5	RIBEIRINHA
6	DOUTRINA	6	EXPEDICIONÁRIO
7	DEFESA	7	ENC
8	COMBATE	8	OPERAÇÃO
9	OPERAÇÃO	9	CCT
10	TROPA	10	ADSUMUS
11	DESEMBARQUE	11	CASC
12	EMPREGO	12	CFN
13	BATALHÃO	13	DESEMBARQUE
14	GPTOPFUZNAV	14	DOUTRINA
15	PRONTIDÃO	15	PRONTIDÃO
16	COMANDANTE	16	ESQUADRA
17	GUERRA	17	EMBARQUE
18	CCT	18	SUBUNIDADE
19	ENC	19	FORÇA
20	MEIOS	20	COMBATE
21	EMBARQUE	21	BATALHÃO
22	ADSUMUS	22	PREPARO
23	CASC	23	TROPA
24	CCA	24	DEFESA
25	CFN	25	EMPREGO
26	CIAB	26	GUERRA
27	EXÉRCITO	27	COMANDANTE
28	MISSÃO	28	SEGURANÇA
29	SUBUNIDADE		
30	EMPREGAR		
31	PIRATARIA		
32	TERRORISMO		
33	PREPARO		
34	SEGURANÇA		
35	GUERRAS		
36	AMEAÇAS		
37	BATALHÕES		
38	PELOTÕES		

Apêndice 5 – Lista de unigramas extraídos com as ferramentas em português

UNIGRAMAS ANTCONC_EN		UNIGRAMAS TERMOSTAT_EN	
1	MARINE	1	OPERATION
2	NAVAL	2	DEFENSE
3	AMPHIBIOUS	3	LANDING
4	BN	4	MARINE
5	COMBAT	5	MATERIEL
6	EXPEDITIONARY	6	WARFARE
7	DEFENSE		
8	RIVERINE		
9	RU		
10	DOCTRINE		
11	OPERATION		
12	WARFARE		
13	LANDING		
14	MILITARY		
15	SECURITY		

Apêndice 6 – Lista de unigramas extraídos com as ferramentas em inglês

BI/TRI/QUADRIGRAMAS ANTCONC_PT		BI/TRI/QUADRIGRAMAS TERMOSTAT_PT	
1	CORPO DE FUZILEIROS NAVAIS	1	FUZILEIROS NAVAIS
2	MARINHA DO BRASIL	2	AÇÃO CÍVICO-SOCIAL
3	OPERAÇÕES ESPECIAIS	3	AMAZÔNIA AZUL
4	OPERAÇÕES RIBEIRINHAS	4	CARÁTER ANFÍBIO
5	DEFESA NBQR	5	CARÁTER ANFÍBIO
6	DOUTRINA MILITAR NAVAL	6	CIA DE POLÍCIA
7	PATRULHA NAVAL	7	COMANDANTE DA MARINHA
8	AMAZÔNIA AZUL	8	CORPO DE FUZILEIRO
9	DIVISÃO ANFÍBIA	9	CORPO DE FUZILEIRO
10	FUZILEIRO NAVAL	10	DEFESA NBQR
11	ASSISTÊNCIA HUMANITÁRIA	11	DIRETORIA DE HIDROGRAFIA
12	COMANDO E CONTROLE	12	DIVISÃO ANFÍBIA
13	EMPREGO LIMITADO DA FORÇA	13	EMBARCAÇÃO DE GUERRA
14	TROPA DE DESEMBARQUE	14	FORÇA AERONAVAL
15	TROPA DE REFORÇO	15	FUZILEIRO NAVAL
16	ÁGUAS JURISDICIONAIS BRASILEIRAS	16	GRUPO DE ASSESSORAMENTO
		17	GUERRA CIBERNÉTICA
		18	GUERRA HÍBRIDA
		19	MARINHA DO BRASIL
		20	MARINHA DO BRASIL
		21	OPERAÇÃO ANFÍBIA
		22	OPERAÇÃO DE GUERRA NAVAL
		23	OPERAÇÃO DE INFORMAÇÃO
		24	OPERAÇÃO DE PAZ
		25	OPERAÇÕES ESPECIAIS
		26	OPERAÇÕES RIBEIRINHAS
		27	PRONTO EMPREGO
		28	TROPA DE DESEMBARQUE
		29	TROPA DE EMBARQUE
		30	TROPA DE REFORÇO

Apêndice 7 – Lista de bi/tri/quadrigramas em português extraídos com as ferramentas

BIG/TRI/QUADRIGRAMAS ANTCONC_EN		BIG/TRI/QUADRIGRAMAS TERMOSTAT_EN	
1	MARINE CORPS	1	PLACE OF STRATEGIC INTEREST
2	BRAZILIAN NAVY	2	RIVERINE OPERATION
3	AMPHIBIOUS DIVISION	3	CIVIC-SOCIAL ACTION
4	NAVAL PATROLLING	4	INTERAGENCY OPERATION
5	BRAZILIAN JURISDICTIONAL WATERS	5	AMPHIBIOUS AND EXPEDITIONARY FORCE
6	MARINE ADVISORY GROUP	6	WAR VESSEL
7	MARINE AMPHIBIOUS BRIGADE	7	HYBRID WARFARE
8	NAVAL SPECIAL OPERATIONS COMMAND	8	HUMANITARIAN OPERATION
		9	LIMITED USE OF FORCE
		10	JOINT OPERATION

Apêndice 8 – Lista de bi/tri/quadrigramas em inglês extraídos com as ferramentas