Tese de Doutorado

PROPOSTA DE ORDEM SEQUENCIAL E CRIAÇÃO DE SISTEMAS INFORMÁTICOS PARA EXTRAÇÃO TERMINOLÓGICA BILÍNGUE EM CORPORA PARALELOS – INGLÊS/PORTUGUÊS – COM VISTAS À TRADUÇÃO DE TEXTOS DAS CIÊNCIAS MÉDICAS

Lautenai Antonio Bartholamei Junior

Universidade Federal de Santa Catarina
Programa de Pós-Graduação em
Estudos da Tradução

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE COMUNICAÇÃO E EXPRESSÃO
PÓS-GRADUAÇÃO EM ESTUDOS DA TRADUÇÃO
Lautenai Antonio Bartholamei Junior

PROPOSTA DE ORDEM SEQUENCIAL E CRIAÇÃO DE SISTEMAS INFORMÁTICOS PARA EXTRAÇÃO TERMINOLÓGICA BILÍNGUE EM CORPORA PARALELOS – INGLÊS/PORTUGUÊS – COM VISTAS À TRADUÇÃO DE TEXTO DAS CIÊNCIAS MÉDICAS

Florianópolis
2013

Ficha de identificação da obra elaborada pelo autor
através do Programa de Geração Automática da Biblioteca Universitária
da UFSC.

A ficha de identificação é elaborada pelo próprio autor
Maiores informações em:
http://portalbu.ufsc.br/ficha

Lautenai Antonio Bartholamei Junior

PROPOSTA DE ORDEM SEQUENCIAL E CRIAÇÃO DE SISTEMAS INFORMÁTICOS PARA EXTRAÇÃO TERMINOLÓGICA BILÍNGUE EM CORPORA PARALELOS – INGLÊS/PORTUGUÊS – COM VISTAS À TRADUÇÃO DE TEXTOS DAS CIÊNCIAS MÉDICAS

Tese submetida ao Programa de Pós-Graduação em Estudos da Tradução da Universidade Federal de Santa Catarina para a obtenção do Grau de Doutor em Estudos da Tradução.

Orientador: Prof. Dr. Ronaldo Lima
Coorientador: Prof. Dr. Alain-Philippe Durand

Florianópolis
2013

Lautenai Antonio Bartholamei Junior

PROPOSTA DE ORDEM SEQUENCIAL E CRIAÇÃO DE SISTEMAS INFORMÁTICOS PARA EXTRAÇÃO TERMINOLÓGICA BILÍNGUE EM CORPORA PARALELOS – INGLÊS/PORTUGUÊS – COM VISTAS À TRADUÇÃO DE TEXTOS DAS CIÊNCIAS MÉDICAS

Esta Tese foi julgada adequada para obtenção do Título de Doutor em Estudos da Tradução e aprovada em sua forma final pelo Programa de Pós-Graduação em Estudos da Tradução da Universidade Federal de Santa Catarina.

Florianópolis, quarta-feira, 27 de novembro de 2013.

_____
Prof.ª Andréia Guerini, Dra.
Coordenadora do Curso

Banca Examinadora:

_____
Prof. Ronaldo Lima, Dr.
Orientador
Universidade Federal de Santa Catarina

_____
Prof. Alain-Philippe Durand, Dr.
Coorientador
University of Arizona

_____
Prof. Antonio Paulo Berber Sardinha, Dr.
Pontifícia Universidade Católica de São Paulo

_____
Prof. Anderson da Costa, Dr.
Universidade Federal de Santa Catarina

_____
Profa. Maria Jose Bocorny Finatto, Dra.
Universidade Federal do Rio Grande do Sul

_____
Profa. Adja Balbino Barbieri Durão, Dra.
Universidade Federal de Santa Catarina

_____
Profa. Cristiana Vieira, Dra.
École Supérieure de Mines de Sainte Etienne

AGRADECIMENTOS

Aquilo que a terminologia romântica chamava de gênio ou talento ou inspiração ou intuição não é outra coisa que o encontrar a estrada empiricamente, pelo faro, cortando por atalhos, lá onde a máquina seguiria um caminho sistemático e consciencioso, ainda que velocíssimo e simultaneamente plural.
(Italo Calvino, 1969)

RESUMO

A extração terminológica bilíngue cada vez mais se firma como um campo de pesquisa explorado por pesquisadores no âmbito dos Estudos da Tradução. Parte considerável das investigações atualmente desenvolvidas volta-se à operacionalização das tarefas de extração terminológica por meio de ferramentas computacionais, produzindo glossários para servirem como ferramenta de apoio aos tradutores. Esta pesquisa de doutoramento desenvolve uma proposta sequencial para a extração terminológica na área das Ciências Médicas, centrando-se principalmente em uma lacuna detectada em estudos anteriores, a saber: a busca por correspondentes tradutórios dos candidatos a termos, geralmente realizada, de forma manual. Nesta perspectiva, o estudo emprega ferramentas fornecidas pelo Processamento da Linguagem Natural (PLN), evocando principalmente as seguintes disciplinas: Linguística de Corpus, Corpora nos Estudos da Tradução, Terminologia e Extração Terminológica, com o intuito de oferecer um processo sistemático que contemple o processo de extração terminológica. Na presente proposta, os dados obtidos evidenciam altos níveis de precisão, que levam a supor que por meio da referida abordagem a recuperação de candidatos a termos e a busca por seus correspondentes tradutórios pode efetivamente ser otimizadas, revelando-se tão eficiente quanto a extração terminológica realizada analogicamente por especialistas. Em uma escala numérica de 0 (zero) a 1 (um), a probabilidade de 0,822645962, 0.969518 e, em alguns casos, 1 (um), explicita a precisão dos correspondentes tradutórios. Os resultados ainda demonstraram que, embora os corpora utilizados para análise estejam expostos em português europeu, e circunscritos uma área especifica do conhecimento, os valores semânticos dos correspondentes tradutórios foram mantido. Na proposta de ordem sequencial e criação de sistemas informáticos para extração terminológica bilíngue em corpora paralelos, a ordem sequencial proposta, tal como os sistemas informáticos desenvolvidos para o processamento dos dados tratam do par de idiomas inglês-português, no entanto, poderão ser utilizados outros pares de línguas e corpora de outros campos do conhecimento.

Palavras-chave: Estudos da tradução. Linguística de corpus. Extração terminológica.

ABSTRACT

Bilingual terminology extraction increasingly firm as a field of research explored by scholars in the context of Translation Studies. Considerable part of the researches currently carried back to the operationalization of terminology extraction tasks by using computational tools, producing glossaries to serve as a support tool for translators. The research developed in this PhD dissertation aims to develop a sequential proposal for terminology extraction in the field of Medical Sciences focusing mainly on a gap detected in previous studies, *viz.*, searching for matching translation equivalents for terms candidates generally held manually. In this perspective, the study uses tools provided by Natural Language Processing (NLP), mostly evoking the following disciplines: Corpus Linguistics, Corpora in Translation Studies, Terminology and Terminology Extraction, in order to offer a systematic process that addresses the terminology extraction task. In the proposal, data gathered presented high levels of accuracy, leading us to believe that through this approach for retrieval of translation equivalents for the terms candidates can be optimized effectively, preserving as efficient in terminology extraction as performed by specialists. In a numerical scale from 0 (zero) to 1 (one), probabilities as in 0.822645962 , 0.969518 , and in some cases 1 (one) explicit translation equivalents retrieval accuracy. The results also showed that, although the corpora used for analysis are exposed in European Portuguese, and circumscribed a specific area of knowledge, the semantic value of the translation equivalents was maintained. In the sequential order proposal and creation of systems for bilingual terminology extraction in parallel corpora , the sequential order proposed, as well the computational systems developed for data processing dealing with the English-Portuguese language pair, however, it cloud be used to other languages pairs and corpora to other fields of knowledge.

Keywords: Translation studies. Corpus linguistics. Terminology extraction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

Estudos da Tradução e Corpora (ETC)

Linguística de Corpus (LC)

Terminologia (TER)

Extração Terminológica (ET)

SUMÁRIO

0. AVANT-PROPOS DO AUTOR

Os tradutores que recorrem ao uso de suportes informáticos como complemento ao seu trabalho, ou mesmo como substitutos a determinadas técnicas análogas de análise e tratamento de dados dispostos em papel, em não sendo especialistas em informática, muitas vezes parecem preferir empregar programas oferecidos ao grande público, sem necessariamente se cercar de preocupações no sentido de mergulhar nos complexos meandros de natureza técnica subjacentes à sua utilização.

Parece natural o fato de que o tradutor, profissional das ciências da linguagem, não esteja sempre interessado em se aprofundar nas tarefas de programação e/ou manipulações; aliás, demasiadamente complexas quando se trata de processar dados da linguagem natural de forma tecnicamente mais profunda. Tampouco parece ser pertinente aos profissionais das letras se especializarem no domínio de comandos sofisticados, que derivam seus focos às ciências exatas, como por exemplo, as ciências da computação. De fato, não há como contornar a premissa de que o uso das novas tecnologias – em seus finos meandros – muitas vezes exige empenho técnico, pois basta que as exigências em termos de processamento se intensifiquem para que os suportes oferecidos ao grande público não satisfaçam determinadas necessidades do tradutor.

Em relação àqueles profissionais cuja formação se constitui de forma híbrida, ou seja, pluri e multidisciplinar, a questão se apresenta de outra forma, isto é, o prolongamento técnico não constitui um problema; pelo contrário, pode trazer soluções. Parece não ser interessante aos tradutores – não especialistas em informática – assumir ônus que implicariam centenas de horas a mergulhar em aperfeiçoamentos, testes e aplicação de técnicas específicas, em geral apanágios do campo matemático.

Enquanto profissional ligado ao universo das ciências da linguagem, insiste-se, todavia, que possa ser sensato supor que os aportes para o refinamento tecnológico dos aparatos destinado ao processamento das línguas, sobretudo desenvolvidos com vistas à tradução, devam contar com a participação de linguistas, tradutores e intérpretes ligados ao estudo do texto em sentido amplo. Logo, atualmente, algum grau de conhecimento a respeito de informática se

faz necessário, mesmo que se julgue, em princípio, ser totalmente possível contornar a adoção de técnicas mais apuradas para o estudo de fatos de natureza linguística. Referimo-nos em particular aos estudos terminológicos ou lexicográficos. O advento das bases ou de dicionários eletrônicos poderia ser um exemplo da referida tendência. O terminólogo e o lexicógrafo parecem ser os profissionais mais habilitados para promover melhores graus de conviviabilidade para o manuseio dos dados que propõem por parte dos usuários.

Entendendo que nem todo intérprete, tradutor ou linguista está preparado para lidar com manipulações informáticas, a presente proposta de estudo visa justamente suavizar o contato do tradutor com parte das restrições e implicações de ordem técnica que se apresentam àqueles que se lançam no campo de processamento de textos. Ao propor uma ordem linear para o uso de programas computacionais com vistas à extração terminológica bilíngue em corpora paralelos dedicados à exploração de textos medicais, pretende-se oferecer algumas contribuições neste sentido.

De forma prudentemente delimitada, busca-se realizar o processo de exploração terminológica a partir de textos circunscritos no campo das ciências médicas. O intuito central é o que buscar expor a viabilidade do emprego de procedimentos automatizados para o processamento de dados de natureza verbal, escrita. Em outras palavras, segundo Yuste Frías (2010) nós, tradutores, não lidamos com as línguas, mas com textos. Como observou Saussure (1969/1916) as línguas se compõem de abstrações. Logo, o que se processa, com efeito, não são as "línguas" – entidades virtuais – mas "textos" tipografados ou orais. São justamente os textos que cristalizam fatias dos discursos e que concedem significações (locais) e sentido (geral) ao material textual. Os resultados dos processamentos dos quais se tratam aqui, em não sendo analíticos em sua essência, mas prioritariamente quantitativos, oferecerão – acredita-se – certa precisão probabilística, ou seja, representam dados eventualmente auxiliares às tomadas de decisões, sobretudo no que concerne à tradução de textos de cunho técnico (ou científico).

Se, por um lado, se lida com a necessidade de manipular sistemas informatizados, de formalizar dados e submetê-los ao processamento automático; por outro lado, os resultados obtidos por meio dos procedimentos aplicados poderão refletir os benefícios da otimização do tempo de trabalho empregado para fazê-lo, ou seja, a disponibilidade que se tem para estudo do material a ser traduzido em razão, entre outros, da profusão excessiva de informações veiculadas no

cenário moderno, em constante e incessante metamorfose exige auxílio da máquina. Em outras palavras, os dados gerados automaticamente representam amostras extraídas de cernes textuais representativos de determinado setor da ciência, selecionados entre milhões de dados. Tal seleção, supõe-se, permitiria, quase sempre, atingir maiores graus de certeza para a tomada de decisões em termos de tradução.

Atualmente, o máximo que se consegue traduzir de forma totalmente automática são textos cujas variações de ordem linguística, semântica, pragmática e conceitual, bem como seus microuniversos referenciais, são totalmente previsíveis (Arnold et al., 1993). Deve-se aceitar, porém, que os esforços no campo técnico-linguístico, no sentido de agregar ferramentas de suporte ao processamento de textos, têm gerado, segundo Hutchins (1992), resultados positivos para a área. É provável que ainda estejamos longe de ver nascerem sistemas de tradução automática capazes de traduzir textos literários abertos, permeados por metáforas, ambiguidades, polissemias e pressuposições. Talvez mais longe ainda de conhecer programas voltados à tradução de poesias não somente de uma língua para a outra, mas também de forma intralinguística. Para tal, basta considerar a métrica, a rima ou traços de natureza suprassegmental, que ultrapassam os domínios da letra (forma).

Circunscritos no domínio digital disponíveis no mercado, não se pode negar a riqueza dos dicionários digitais, dos glossários, das enciclopédias, dos tradutores semiautomáticos, dos compiladores, dos sistemas de localização de sequências gráficas disponíveis atualmente. Enfim, há uma quantidade considerável de instrumentos de suporte que podem constituir diferenciais para os tradutores modernos, isto é, àqueles que buscam aperfeiçoar ações apoiadas em suportes confiáveis, mesmo ao lidar com grandes volumes de informação (Archer, 2002).

A necessidade de se traduzir textos de campos de conhecimento especializados se eleva paralelamente à demanda por materiais de suporte à tradução. Tendo em vista os imensos volumes de dados a serem traduzidos diariamente, em curtos espaços de tempo, não se pode negar que há um grande movimento – bastante importante aliás – no desenvolvimento de ferramentas para processamento e análises pontuais. Um dos exemplos a ser particularmente citado, concerne ao *Extrator de Legendas de Filmes* (Bartholamei, 2011), elaborado por este pesquisador em 2011, precisamente para responder a demandas prementes de estudiosos que, anteriormente, dispensavam dezenas de horas copiando textos manualmente de mídias visuais, por vezes difíceis de serem manuseadas. Com o auxílio do referido programa além da extração automática é possível *atrelar* a expressão imagética ao material

linguístico correspondente. Em outras palavras, torna-se possível localizar frases, palavras e símbolos a serem instantaneamente requeridos juntamente com as respectivas imagens às quais o texto se refere. Posteriormente, sobre o texto extraído, todas as outras opções de processamento automático obedecerão aos mesmos princípios disponibilizados pela Linguística de Corpus. A multidisciplinaridade que abarca o processamento, estende-se ao visionamento[1] de mídias digitais de forma ampla, gerando grande progresso para os estudiosos dessa área de estudos relativamente recente.

De fato, o tratamento de grandes volumes textuais parece se tratar de tarefa difícil de ser levada a cabo em curtos espaços de tempo por meio de ação intelectual humana. Logo, um dos objetivos paralelos à presente proposta, é justamente o pôr em evidência de que, se há alguns anos tais operações eram apanágio de discursos de especialistas em informática, ou tarefas eventualmente julgadas utópicas; atualmente o caráter convivial dos sistemas de processamento, aliado a suas grandes potencialidades, vêm contemplando público cada vez mais amplo e exigente. Tais avanços vêm lenta e progressivamente permitindo que os utilizadores se libertem de eventuais preocupações com entraves de ordem técnica, subjacentes aos naturais conflitos em relação ao rápido avanço dos aparatos tecnológicos, e passem a acreditar na possibilidade de domínio da linguagem informática em prol dos estudos da linguagem natural. Os atuais formatos dos programas, bem como os percursos aqui apresentados visam "desobstruir" alguns encalços e restrições que, outrora, tornavam o processamento do texto algo reservado a especialistas com conhecimentos atestados no ramo computacional. Doravante, acredita-se que muitos intérpretes e tradutores se disponham a gerar seus próprios aportes *à la carte*, ou seja, moldados à aura da ergonomia calculada.

O elevado grau de desempenho dos programas atualmente disponíveis no mercado exige, no entanto, esclarecimentos e precisões ao profissional das letras que pretenda considerar o resultado de cálculos probabilísticos para respaldar suas decisões no tratamento do texto. De fato, não se pretende, absolutamente, menosprezar a possibilidade de geração de dados por meio dos estudos análogos e qualitativos, realizados a partir do discernimento e da razão, sobretudo levando-se em consideração que a riqueza e precisão das análises humanas dificilmente

[1] Visionar mídias digitais tornou-se um campo em franco progresso a partir dos anos 1980, quando o advento da microinformática permitiu, a qualquer indivíduo interessado, lançar-se na produção e análise de vídeos: legendando, transformando, produzindo.

poderão ser substituídas pela máquina, mas tão somente vislumbrar e a possibilidade de que o aperfeiçoamento das técnicas pode gerar progressos importantes para o processamento e consequente obtenção de dados cada vez mais confiáveis. Como se aventou acima, dificilmente as máquinas estarão preparadas, nas próximas décadas, para enfrentar fenômenos linguísticos sobre os quais não há acordos definidos, como é o caso da ambiguidade, dos pressupostos, ou simplesmente dos relativos.

Acredita-se, particularmente, que muitos progressos em processamento automático poderiam e deveriam ser realizados por profissionais das letras, pois em conhecendo especificidades das línguas, provavelmente estariam mais habilitados para unir o conhecimento sobre as línguas com a *expertise* dos especialistas que trabalham na concepção de produtos informáticos destinadas ao tratamento de dados de natureza linguística.

Os excessos dos imaginários criados e, algumas vezes, lançados sobre os supostos malefícios do digital frente ao analógico, por muito tempo geraram críticas ferrenhas que encontram eco na carência do qualitativo face ao quantitativo ou, similarmente, da excelência do trabalho humano frente à frieza dos resultados gerados pela máquina. Sabe-se que o estudo de uma única entidade lexical, examinada em seus contextos discursivos pode gerar muitas teses. De modo similar, o confronto de uma única unidade lexical examinada no âmbito de um par de línguas pode conduzir a muitos tratados antropológicos, sociológicos, políticos, culturais e históricos. Porque então insistir no uso de suportes informáticos para o tratamento de dados de natureza linguística se a máquina ignora todas estas implicações?

De fato, a resposta não é simples, tampouco haveria uma única explicação, pois enquanto pesquisadores estamos todos situados no cerne das ebulições e das mudanças em progresso, sem poder ainda determinar com exatidão onde tais progressos poderão desembocar. Todavia, nos dias atuais sabe-se que instrumentos como o *Google Tradutor*, ou os *dicionários on-line*, os *glossários informatizados* ou, ainda, a própria *Internet* de forma geral, empregados para o cálculo de ocorrências, têm sensibilizado tradutores, linguistas e estudiosos autônomos que encontram nestas ferramentas pistas para aceitar, recusar ou até mesmo para anular suas hipóteses primárias. Logo, toma-se aqui o processamento não como um fim certo e concreto, mas, sobretudo como um meio auxiliar no âmbito do processo de tradução, capaz de realizar de forma bastante satisfatória certas tarefas.

O que se propõe no presente trabalho parece se situar nos espaços limítrofes entre o qualitativo e o quantitativo, posto que as sequências de uso de programas, bem como o preenchimento de lacunas através da criação de novos instrumentos derivam para o campo da Inteligência Artificial rudimentar, mesmo que de forma prioritariamente numérica e não analítico-valorativa. De fato, atualmente não há como se falar em avaliação de dados linguísticos de forma inteligente por meio do uso de computadores. Todavia, é importante considerar que os aperfeiçoamentos informáticos paulatinamente ultrapassam a simples análise estatística. A *lexicometria* atual, por exemplo, não aponta somente para sequências de caracteres em termos de ordem de frequência, mas é também capaz de garantir a instauração de relações confiáveis entre materiais lexicais, atrelando as ocorrências aos diversos contextos em que emergem.

Entende-se aqui por *lexical* não somente a palavra enquanto forma, mas também enquanto estrutura dotada de *significação* (isolada) e de *sentido* (em contexto). As conversões dos dados linguísticos análogos para o formato binário, com base nas propostas que serão apresentadas ao longo deste estudo, não eliminam, absolutamente, a excelência das informações semânticas, pragmáticas e conceituais inerentes à expressão, ou seja, ao **verbo em contexto**, elevado ao patamar de discurso e, em última instância, de linguagem; pelo contrário, buscam, sim, conservar traços inerentes às composições de sentido através da consideração dos contextos discursivos, sobretudo diante do fato de que os problemas de limitações para armazenamento de dados já não constituem mais obstáculos ao processamento de dados linguísticos, sejam eles de que natureza for: escrito, oral, espectral ou codificados. A capacidade de armazenagem de informações responde a praticamente todas as exigências atuais e mesmo daquelas projetadas, por exemplo, para os próximos 10 anos.

De modo breve e conclusivo, a preocupações em relação à suposta invasão do numérico nas atividades sociais não se asseveram verdadeiras. A experiência do tradutor humano cada vez mais reafirma seu papel, colocando em segundo plano quaisquer instrumentos que possam ser agregados a seu trabalho. Deste modo, embora se insista aqui, sobre as qualidades e benefícios do processamento de textos, tem-se plena consciência de que se trata tão somente de um campo de investigação que oferece ferramentas e técnicas em prol da extensão e da agilização das atividades humanas ligadas ao tratamento de dados linguísticos. Neste sentido, negar as contribuições da Linguística de

Corpus para os Estudos da Interpretação e da Tradução não traz nenhum benefício para a área.

1 INTRODUÇÃO

1.1 ORGANIZAÇÃO DA TESE

O presente texto de Tese está dividido em oito seções, a saber:

Introdução à pesquisa;
Fundamentação teórica;
Fundamentação metodológica;
Metodologia;
Análise dos dados;
Proposta sequencial;
Considerações finais;
Referências.

Na primeira seção, apresentam-se algumas reflexões do autor sobre implicações das investigações voltadas ao processamento linguístico no âmbito dos estudos das línguas de modo geral, e da tradução de forma específica. Destaca-se a relevância do tema proposto para estudo, bem como se busca evidenciar as eventuais discussões e aportes que a investigação evoca, tanto para os estudos de corpus com vistas à tradução de textos de conteúdo técnico, como às atividades ligadas ao trabalho prático do tradutor. A seção introdutória avança informações sobre a organização da tese e, por fim, apresenta uma projeção estrutural de como será desenvolvido texto final deste estudo.

Na segunda seção, Fundamentação Teórica: ETC, LC, TER e ET, apresenta-se um panorama dos suportes teóricos imediatos à pesquisa, bem como a Revisão da Literatura, referenciada como base para o desenvolvimento da presente pesquisa. Conceitos referentes aos Estudos da Tradução e Corpora, Linguística de Corpus, Terminologia e Extração Terminológica serão abordados e discutidos em consonância com a proposta do trabalho desenvolvido, tendo em vista suas extensões amplas. Menções a teóricos, estudiosos e pesquisadores serão trazidos à baila, de forma que se seja possível estabelecer relações de complementaridade entre as diversas literaturas apresentadas e seus postulados metodológicos decorrentes. A intenção é evocar subsídios preconizados e compartilhados na área da linguística de corpus, mantendo os desenvolvimentos circunscritos neste escopo, e entre especialistas.

As delimitações e definições parecem se firmar como guarda-chuva metodológico com vistas a garantir a concatenação científica, funcionando como base de partida para construção dos saberes caracterizados no presente campo. Entre os principais conceitos discutidos, destacam-se:

a) o uso de corpora nas pesquisas em terminologia e extração terminológica com vistas aos estudos e prática da tradução;

b) a integração e diálogo entre aparatos informáticos da linguística de corpus e sua ligação com os estudos da tradução, principalmente nas abordagens que admitem o uso de corpora como suporte às práticas de interpretação e de tradução – e não somente voltado à composição da crítica ou da teorização;

c) a discussão sobre terminologia e tradução;

d) e finalmente, questões referentes às ações dos principais precursores e especialistas da extração terminológica.

Na terceira seção, Metodologia: Projeto, Construção e Processamento do Corpus, explicitam-se as etapas adotadas para a execução da pesquisa. Inicialmente, apresenta-se o planejamento para a definição do corpus de estudo, destacando suas características inerentes e propósitos de sua adoção. Em seguida, apresenta-se o processo de preparação do corpus desde sua apreensão até seus níveis mais complexos de codificação, anotação e alinhamento. Por último, a etapa de processamento do corpus é apresentada. Para fazê-lo, expõe-se ao leitor de que forma o conjunto de materiais textuais tratados e reunidos para determinado fim será processado, no sentido da obtenção de dados para posterior análise eventual consulta.

A quarta seção, intitulada: Dados: Análise, trata de discussões endereçadas à proposta de uma ordem sequencial para a geração de glossários bilíngues, como também a análises dos dados iniciais obtidos a partir dos corpora processados. Por fim, discutem-se implicações decorrentes dos resultados automaticamente obtidos por meio do processo de extração terminológica adotado, focando principalmente a busca automática por correspondentes tradutórios em corpora bilíngues paralelos. Encerra-se assim esta parte.

A sexta seção apresenta o desenho da proposta de ordem sequencial e criação de sistemas informáticos para extração

terminológica bilíngue em corpora paralelos – inglês/português – com vistas à tradução de textos das ciências médicas.

Na sétima seção, intitulada: Considerações Finais, retoma-se algumas das questões centrais levantadas durante o processo de pesquisa e redação. Com base nos objetivos, resultados e análise dos dados, busca-se uma visão que permita defender a pertinência de um trabalho geralmente atribuído a programadores, engenheiros e técnicos em informática, mas cujo foco recai sobre textos e, por extensão, sobre as línguas. Logo, reascende-se a dúvida sobre a necessidade de participação do profissional com formação híbrida: em informática e ciências da linguagem.

Referências Bibliográficas, a oitava e última seção expõe os trabalhos citados na composição do presente texto, bem como instrução para os procedimentos teóricos e metodológicos.

1.2 APRESENTAÇÃO

O estudo e reflexão sobre a crítica, a teoria e a prática em tradução no âmbito acadêmico-científico é considerado por pesquisadores como Holmes (1972/1988) como sendo um dos marcos do nascimento de um novo campo de estudos, a saber: a disciplina ***Estudos da Tradução***. Em seu texto, considerado como seminal no processo de instauração da nova área, apresentado no artigo intitulado: "*The Name and Nature of Translation Studies*" (O Nome e a Natureza dos Estudos da Tradução), James Holmes discute fundamentos não somente para fortalecer a instauração da disciplina em questão, como também identifica e delineia parte considerável dos tópicos abordados no então, "novo campo", que nascia no momento em que lançou seu trabalho. Na ocasião, o referido autor propôs uma série de tópicos que caracterizavam, efetivamente, os estudos da tradução como uma *interdisciplina* (cf. Delisle, 1998). Em seu mapa inicial já é possível constatar os raios de ação dos Estudos da Tradução.

Entre as subáreas ligadas à tradução, muitas se consolidaram como permanentes e vão ao encontro da proposta aqui apresentada. Referimo-nos especificamente ao título deste trabalho, que aponta para a proposição de aparatos e meios técnicos com vistas à constituição de suportes informáticos orientados ao processamento automático de textos. Mais precisamente, de forma clara e explícita, evidenciam-se informações sobre uma sequência de procedimentos científicos em sintonia e em diálogo, metaforicamente *simbiótico*, entre os diversos

sistemas que compõem o pacote proposto; de forma não somente linear e unilateral, mas também retroativa, com instrumentos passíveis de serem empregados como mecanismos com funções duais, ou seja: (a) para o processamento de dados que, por sua vez, se tornarão (b) fontes de informações ao tradutor. Com efeito, trata-se de modelos estatísticos para a extração de termos em duas línguas: português-inglês, a partir de corpora paralelos, constituindo-se em via inovadora, concebida *à la carte*, visando o aperfeiçoamento de processos de busca por dados linguísticos confiáveis do ponto de vista quantitativo particularmente dedicados à área médica.

Na condição de quem, assumidamente, não pretende questionar produtos e processos já atestados cientificamente (leia-se *reinventar a roda*) por um lado, busca-se manter a proposta de trabalho circunscrita nos campos da disciplina de Estudos da Tradução apresentados por Holmes (1972/1988) que, por sua vez, preconiza a construção de plataformas focadas na automatização de processos de extração de dados de natureza textual e, por conseguinte, a geração de glossários bilíngues por meio de processos estatísticos. Por outro lado, naturalmente, a condição de pesquisador exigiu que se procurasse garantir alguns *passos adiante* em prol dos estudos científicos na área em questão. Neste sentido, humildemente, derivam-se intenções à elaboração de novos sistemas integráveis, em âmbito de concepção pessoal deste pesquisador, às bases método-epistemológicas de uma área já consolidada e, atualmente, integrada aos Estudos da Tradução, ou seja, a Linguística de Corpus.

O interesse pelo estudo, como assinalado no *Avant-Propos* desta tese, acima postulado, deve-se principalmente ao advento da introdução e plena aceitação da microinformática na execução das mais diversas atividades humanas, tanto em âmbito cotidiano e pessoal, quanto profissional. A velocidade em que circulam as informações nos dias atuais exige, em alguns setores nos quais grandes volumes de textos são processados, que se recorra a instrumentos de suporte para apreensão e processamento. Parece que poucos tradutores não empregam, nos dias atuais, algum tipo de material informatizado. No limite, sobressai o imperceptível, o consubstancial. Por exemplo, não mais datilografamos, mas digitamos, trabalhando sobre a conversão do análogo à ordem binária subjacente aos dados posteriormente impressos. Todavia, o utilizador (grande público) não possui, necessariamente, consciência dos processos que subjazem os dados expostos na tela.

Acrescenta-se ainda o fato de que, dependendo do tipo de enquadramento do tradutor, suas atividades recaem sobre materiais de

diversas áreas do conhecimento. Ora, textos específicos se caracterizam por utilizar as estruturas e sentidos linguísticos próprios à língua sobre a qual se erguem (dita *língua geral*, ou *de base*), porém as línguas de especialidade, aquelas de domínio específico, se caracterizam por tendências gráficas, lexicais, semânticas, pragmáticas e conceituais que as distinguem da língua de referência. O tradutor profissional, diante de textos de especialidade, não difere do tradutor literário em muitos sentidos. Por exemplo, ambos precisarão dispor de grandes quantidades de informações intra e extratextuais que permitam atribuir sentido tanto ao texto de partida, quanto ao texto de chegada. Finalmente, o caráter científico das línguas de especialidade exige ancoragem pragmática para sua interpretação e representação em outro código.

Acredita-se que os trabalhos com corpora possam minimizar algumas implicações que, por vezes, tornam a tradução científica bastante penosa e demorada, tal como a definição dos candidatos a termo, foco desta pesquisa. Com efeito, a máquina dificilmente conseguirá superar a razão humana, o discernimento, as decisões e os movimentos cognitivos que permitem situar fatos e objetos no tempo e no espaço, como o faz qualquer indivíduo que domine alguma língua ou linguagem expressiva. Todavia, os sistemas informáticos podem ser interessantes àqueles que desejam elucidar fatos linguísticos ao torná-los mais salientes e evidentes por meio de suas ocorrências ou por meio de sua localização automática que se ocultam na complexa trama formada por milhões de letras, símbolos e palavras concatenadas.

McLuhan (1969), nos final dos anos 1960, ao prever e avaliar o futuro impacto que provocaria a introdução de aparatos tecnológicos em praticamente todos os setores da vida cotidiana, atentou para o processo de extensão das capacidades humanas com vistas a acompanhar a velocidade na qual a informação passaria a circular no mundo mecanizado. Carros, aviões, ou mesmo objetos considerados banais como garfos, facas e colheres se tornaram instrumentos que, de certa forma, cibernizaram e estenderam as capacidades humanas. Cálculos de ordem levemente complexa desembocam sobre o uso das calculadoras, cujo valor as tornaram objetos sem valor, mas onipresentes.

Novas abordagens de preparação de glossários bilíngues, por exemplo, passaram a constituir ferramentas de auxílio ao tradutor especializado. Trabalhos desenvolvidos já a partir dos anos 1960 têm por base propostas de extração através de manipulações conduzidas e dirigidas por humano (cf. Hatcher, 1960). No entanto, os avanços atuais, gerados pelo advento da microinformática têm viabilizado o desenvolvimento crescente de aparatos mais leves e compactos,

passíveis de serem rodados em máquinas convencionais de média capacidade. Segundo Bourigault (1992), a agilidade e leveza dos programas têm acontecido de forma cada vez mais orientada ao autogerenciamento, isto é, o utilizador passou a dispor de programas fechados voltados tanto à consulta e leitura, quanto de sistemas abertos concebidos para a realização de acréscimos, inclusão de paratextos (notas, comentários, acréscimos), extrações de excertos *à la carte*, definição de sequências de *macros*, e inclusão de comentários subjacentes para uso do tradutor, entre outras funções criadas *pontualmente* pelo próprio usuário em consonância com suas práticas de trabalho. Tais recursos interessam ao tradutor de textos técnicos.

Em Méndez-Cendón (apud Byrne, 2009) a autora acrescenta que a mera apresentação de informações em um texto científico[2] não basta. A realidade que envolve este gênero textual se vê afetada por uma gama de questões e fatores a considerar, tais como a comunicação técnica, o estilo textual compartilhado e próprio ao domínio científico, a terminologia especializada, o fluxo de trabalho do tradutor, a comunicação multimodal, as exigências legais, a própria tecnologia, e até mesmo a psicologia e a pedagogia envolvidas neste escopo. Com base nestes fatores, apontados por Méndez-Cendón, destaca-se que este trabalho se apropria de dois deles, quais sejam: (a) a terminologia e (b) a tecnologia, ambas noções-chave para a continuidade das investigações em desenvolvimento neste contexto de tese, fortemente centrados na utilização e integração de procedimentos técnicos com vistas à extração terminológica para a geração de instrumentalizações de auxílio ao tradutor, sobretudo em razão das necessidades já apontadas e ligadas à ágil produtividade que marca o cenário dos mercados nos dias atuais.

Baseados nesta perspectiva, situada entre os trabalhos desenvolvidos com interferência humana e as ações realizadas por meio de suportes computacionais automatizados, a presente investigação tem por finalidade a combinação entre modelos manipulados – e manipuláveis –, e aqueles semiautomáticos, ambos visando à extração terminológica bilíngue, no caso específico deste estudo, a partir de corpora paralelos, em consonância com o título desta tese de doutoramento, que define seu fio condutor, tal como especificado com

[2] No escopo do presente trabalho, ora nos referiremos aos texto medicais como técnicos, ora como científicos. Apesar da possibilidade de estratificar os textos em técnicos ou científicos, no caso da área medical, alguns textos revestem-se de maior grau de tecnicidade, outros tratam da ciência. Logo, parece-nos incongruente excluir a quaisquer uma das referidas designações em detrimento da outra.

precisão na explanação do(s) objetivo(s) da atual proposta, exposto abaixo e intimamente ligado ao título desta proposta, que preconiza a sequência e criação de sistemas informativos para a extração terminológica bilíngue em corpora paralelos: inglês/português, com vistas à tradução de textos das ciências médicas.

Considerando os trabalhos já realizados por este doutorando com vistas a oferecer propostas metodológicas para o processamento de dados linguísticos através de ferramentas de compilação e processamento de corpora (cf. Bartholamei Jr, 2009, 2010, 2011), bem como a criação de um glossário técnico bilíngue na área dos Estudos Surdos (cf. Bartholamei Jr, 2008), além de bancos terminológicos como ferramenta de auxílio à tradução (cf. Bartholamei 2009, 2010, 2011), podemos, humildemente, nos considerar aptos à condução dos trabalhos necessário para levar a cabo a proposta e, provavelmente, atingir os objetivos estipulados no escopo da presente pesquisa.

O material linguístico selecionado para a demonstração aplicativa do aparato tecnológico posto em funcionamento, concerne ao campo das ciências médicas. Foram utilizados textos disponíveis no sítio da *EMA* (*European Medicines Agency*), compreendendo a vasta extensão de estudos das ciências médicas, no âmbito da comunidade europeia, respeitando as leis de proteção aos direitos autorais da CEE e as Legislações Internacionais[3].

Os procedimentos metodológicos adotados para este estudo constituirão a base fundamental para o desenvolvimento de ferramentas

[3] Política de direitos autorais:
De acordo com a atual Legislação Internacional e da União Europeia, a Agência Europeia de Medicina (EMA) tem direitos de autoria e outros direitos de propriedade intelectual nos documentos que produz. Documentos classificados como 'Públicos' são disponibilizados a quem lhes interessar e podem ser reproduzidos e/ou distribuídos, no todo ou em parte, independentemente dos meios e/ou os formatos utilizados, para fins não comerciais e também comerciais, desde que EMA seja sempre reconhecida como a fonte do material. Tal reconhecimento deve ser incluído em cada cópia do material. Citações podem ser feitas a partir de tais materiais sem permissão prévia, desde que a fonte seja sempre identificada. A EMA é indenizada por – e contra – todos os custos, processos, reclamações, despesas e passivos decorrentes de quaisquer descumprimentos por qualquer pessoa jurídica ou física, como resultado de qualquer representação ou garantia de ser uma informação errônea. Essas permissões não se aplicam a conteúdos fornecidos por terceiros. Portanto, para documentos onde coletes de direitos autorais de terceiros, a permissão para a reprodução deve ser obtida a partir deste detentor dos direitos autorais. Esta política entra em vigor em 01 de janeiro de 2008. (tradução nossa)
http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000177.jsp

para a extração terminológica, em conformidade com os trabalhos de ponta realizados em grandes centros de pesquisa, tal como: University of Birmingham, Lancaster University, University of Manchester Institute of Science & Technology (UMIST), University of Essex, University of Colorado, University of Edinburgh, Georgetown University, Hong Kong University, University of Nijmegen, University of Victoria. Destacam-se os trabalhos fundamentais de Church e Gale (1991) no desenvolvimento do conhecido algoritmo de alinhamento para a criação da primeira ferramenta de alinhamento estatístico, o *Aligner*, utilizada desde então para a compilação de corpora paralelos e para o desenvolvimento de sistemas de tradução automática. Danielsson e Ridings (1993) apresentaram o *Vanilla Aligner*, com a implantação do algoritmo de Church e Gale (1991) e impulsionaram o desenvolvimento de sistemas com o uso do método estatístico de alinhamento.

Posteriormente, nos laboratórios da University of Twentee, por Hiemstra (1998), foram acrescentados aperfeiçoamentos visando à utilização de ferramentas de alinhamento automático de corpora paralelos ao nível da sentença e da palavra, formando um conjunto semiautomático para o alinhamento estatístico. Na Universidade de Minho houve o desenvolvimento do *NATOOLS* por Simões e João Almeida (2003), que se concentrou em oferecer uma matriz de alinhamento para textos paralelos bilíngues que pode ser tomada como um tipo de abordagem direcionada, sobretudo àqueles que buscam maior entendimento sobre a questão do alinhamento ao nível da sentença e da palavra.

O emprego de uma metodologia baseada em corpora constitui um conceito emprestado à Linguística, principalmente da vertente voltada à linguística de corpus e tem se tornado, como já afirmado, ferramenta importante também nos estudos da interpretação e da tradução, tal como afirma Berber Sardinha:

> Há uma unanimidade entre os pesquisadores da tradução e os linguistas de corpus em torno da questão da utilização de corpora eletrônicos na tradução: o posicionamento corrente é o de que tanto os estudos tradutológicos como área acadêmica de pesquisa, quanto a prática tradutória, têm muito a ganhar com um contato maior com a Linguística de Corpus. (BERBER SARDINHA, 2002, p. 15).

A unanimidade entre os pesquisadores, citada por Berber Sardinha (2002) no excerto acima, tornou notável a utilização de corpora e sua utilização para sua criação e processamento, baseando-se primordialmente no fato de que ao dispor de textos em formato eletrônico (passíveis de processamento por programas computacionais) possibilita-se aos pesquisadores a manipulação e obtenção de dados com maior eficácia e rapidez (cf. McEnery e Wilson, 1996), permitindo acesso a dados até então impossíveis de serem obtidos de modo analógico, isto é, manualmente (Baker, 1993).

Ao perceber a popularidade crescente do uso de corpora nos Estudos da Tradução (Laviosa, 1998), bem como sua aplicação neste campo disciplinar (Bowker, 2002), observa-se que tal abordagem conquistou espaço como recurso à realização de novas frentes de investigação ligadas aos Estudos da Tradução. Pode-se citar como exemplo a realização de pesquisas que exijam o manuseio de grandes volumes de dados e que inviabilizem a ação humana, entre os quais, trabalhos ligados ao processamento de grandes volumes de textos. Salienta-se, ainda, que tal abordagem, como já destacado, desdobra-se em uma série de aplicações que evidenciam usos em contexto, passíveis de interessar ao tradutor.

Investigações sobre corpora têm sido intensificadas nos últimos anos, fazendo com que estudiosos e pesquisadores advoguem em favor da utilização dessa abordagem em diversas áreas e aplicações que exijam suportes estatísticos e/ou probabilísticos. A título de exemplo, pode-se citar, no ensino Hunston (2002); em pesquisa de problemas ligados às estratégias para a tradução Aijmer e Altenberg (2000); e em conformidade com a presente pesquisa para estudos em terminologia Bowker e Pearson (2002) a utilização de corpora na identificação de correspondentes terminológicos em corpora paralelos por, muitas vezes, ser opção mais adequada e eficaz do que a pesquisa nos dicionários tradicionais.

Dentre os procedimentos metodológicos mais comumente adotados pelos pesquisadores em trabalhos de extração terminológica pode-se citar:

a) a construção de corpora tanto monolíngues, quanto bilíngues;

b) a preparação de corpora para seu processamento, envolvendo procedimentos como a verificação de integridade dos materiais

contidos, formatação para o uso com ferramentas de processamento; alinhamento no caso de corpora paralelos;

c) na etapa de processamento dos corpora geralmente se utiliza a comparação entre listas de palavras geradas a partir dos materiais processados para corpora de estudo e corpora de referência, de modo a computar as palavras-chave e, por fim, a busca por correspondentes tradutórios para construção do glossário bilíngue.

Neste estudo, será utilizado um corpus paralelo inglês-português composto por cerca de 30 milhões de palavras do campo das ciências médicas. Textos coletados a partir do portal da *European Medicines Agency*. Será igualmente utilizado um corpus de referência, a saber: o *BNC* (*British National Corpus*) em sua versão completa. Como o *BNC* compõe-se de, aproximadamente, 100 milhões de palavras, acredita-se que se torna representativo da língua inglesa britânica. Os textos, por sua vez, são redigidos e traduzidos no âmbito da comunidade europeia. Observa-se que foi a opção que melhor respondeu às metas da presente pesquisa por se tratar de documentos em situação de uso concreto e atestados pela comunidade científica no qual se enquadram.

Em geral os estudos em terminologia são realizados sobre domínio específicos, ou seja, não se trabalha sobre campos demasiadamente amplos. Por exemplo, em se tratando de Medicina, enquanto área do conhecimento, há alguns anos seria sensato trabalhar, por exemplo, sobre suas subáreas: da Medicina > à Pneumologia > à Alergologia. No caso deste estudo, escolheu-se trabalhar sobre a grande área, Medicina, de forma hiperonímica, em razão da atual possibilidade de processamento de grandes volumes de dados, ou seja, não temos mais, como há 10 ou há 20 anos, limitações em relação à capacidade de processamento e armazenamento de dados. De modo similar, a representatividade dos corpora permite garantir elevado grau de confiabilidade nos resultados alcançados. Como já mencionado na introdução, trata-se do *EMA* (*European Medicines Agency*).

Concomitantemente, como já apontado, busca-se aprofundamentos teóricos em relação às bases informáticas subjacentes aos softwares que, ao final do processo, poderão conduzir a eventuais aprimoramentos dos sistemas no sentido de oferecer buscas mais pontuais e, por conseguinte, resultados mais pertinentes; por um lado, em relação aos correspondentes para as atividades de prática da tradução

e; por outro, como suporte à realização de estudos crítico-teóricos que decorrerão dos resultados alcançados.

1.3 JUSTIFICATIVA

Há, com efeito, a possibilidade de se relacionar três questões. Há, igualmente, a possibilidade de pautá-las como interdependentes, ou seja, as eventuais relações e implicações que as unem, além de inerentes, são inexoráveis no sentido de que não se pode ter controle restrito sobre elas, quais sejam:

(i) desdobramento dos saberes;

(ii) o surgimento de novas disciplinas científicas;

(iii) a adoção de termos e expressões linguísticas que caracterizam as novas *línguas de especialidade*;

(iv) a necessidade por parte do tradutor da terminologia *ad hoc*.

Se, por um lado, a língua de base permeia e define os movimentos que singularizam a comunicação em um campo específico; por outro lado, as fronteiras entre as *Languages for Specific Purposes* (*LSPs*) além de intrinsecamente não serem absolutamente estanques, por vezes geram alterações nos traços semânticos de uma série de componentes lexicais. Tal fenômeno define, aliás, as próprias *LSPs*. Igualmente, novas composições passam a integrar a nova *LSPs* em tal medida que seu processamento para atestação ou é realizado através do domínio dos especialistas da área, ou através do exaustivo trabalho realizado por tradutores. Neste sentido, a linguística de corpus surge como recurso que, naturalmente, pode aperfeiçoar a organização entre o saber e sua expressão, entre a expressão intradisciplinar e a expressão intradisciplinar interlinguística.

Longe de se imaginar uma ilha tecnológica como aquela mencionada por Bioy Casares (1940), ou seja, um universo árido, no qual imperam processos informáticos, a linguística de corpus, em suas evoluções mais recentes, tem procurado considerar identidade, cultura e localização espacial. Ora, se tais aspectos são apanágio à tradução literária, embora reduzidos e minimizados, também emergem, naturalmente em diferente medida na tradução de textos técnicos. Mas

se parcela qualitativa, por conseguinte subjetiva, não pode ser processada facilmente, poderá emergir nos paratextos, manifestados em forma de notas, apontamentos, comentários, remissões e destaques.

Com base em pesquisas realizadas com o intuito de proporcionar alternativa aos tradutores na tarefa de tratar elementos terminológicos oferecendo-os como ferramentas de auxílio, tal como em Coulthard (2005) na tradução científica de textos médicos por meio de seu estudo focado nas traduções do Jornal de Pediatria (JPED); em Souza (2008) investigando a questão lexicográfica do hebraico bíblico, em Siqueira Nobile (2008); em Silva (2008) e, também, Marian (2010), que projetaram seus esforços na investigação de fenômenos lexicográficos na tradução de textos legais, jurídicos e de cunho técnico, como também na comparação de dicionários bilíngues jurídicos para o mercado brasileiro; o conjunto dos trabalhos citados apresenta pontos comuns, constituindo referências para o desenvolvimento de pesquisas no âmbito da extração terminológica bilíngue com vistas a responder necessidades prementes dos Estudos da Tradução.

Com base nos estudos citados e na aspiração pela criação de glossários bilíngues em áreas de especialidade como a da medicina aqui em questão, e com o objetivo de proporcionar ao tradutor uma ferramenta de auxílio por meio da automatização deste processo, atentou-se, então, para o desenvolvimento de uma plataforma que pudesse agregar ferramentas existentes, como, por exemplo, as de alinhamento e o desenvolvimento de novos produtos, como a recuperação de correspondentes tradutores frente os dados obtidos pela mensuração das ligações entre os itens lexicais em corpora paralelos, por meio do alinhamento ao nível de palavra e em conformidade com a relevância de determinados itens lexicais, com base em sua probabilidade para a geração automatizada de glossários bilíngues.

A motivação para o desenvolvimento de um trabalho relacionado à área de extração terminológica, como já observado, parte dos estudos baseados em corpora que, por meio de dados estatísticos, têm servido como fonte para tradutores na solução de questões referentes à terminologia em sua atividade tradutória maior.

Vislumbra-se, igualmente, a integração acadêmico-científica, face à possibilidade em dar continuidade a trabalhos recentes realizados por pesquisadores, tal como observado em Portolan (2011), que se debruçou sobre a elaboração de glossários bilíngues na área de pediatria, cujos esforços derivaram para ampliar o campo do conhecimento da área da medicina e aperfeiçoamento do processamento informático com vistas a propor um conjunto de ferramentas para a realização dos

trâmites de tratamento do léxico em sua integridade em termos de incoativos e conclusivos.

Tomando por base as ferramentas oferecidas pelo campo disciplinar da linguística de corpus e agregando-as em uma proposta concreta, principalmente no que se refere à geração de glossários bilíngues, com interferência humana conduzindo ao desenvolvimento de meios que tornem a produção de glossários bilíngues ferramentas efetivas de auxílio ao tradutor, digam-se *mais eficientes* uma vez que a tradução científica encontra um vasto leque de área de conhecimento, principalmente, a partir dos anos 90, pretende-se disseminar bases desse conhecimento aparentemente restrito a especialistas em informática.

Por fim, a motivação pelo desenvolvimento de projetos e sistemas de extração terminológica baseia-se na possibilidade de realização de buscas por correspondentes tradutores e geração de glossários bilíngues em diversos idiomas e áreas de interesses, não se tornando específico para determinado par linguístico. Sendo assim, como ferramenta de auxílio ao tradutor, a proposta de ordem sequencial e suas ferramentas atreladas, poderão ser amplamente usadas e exploradas tanto por tradutores profissionais, quanto por estudantes, pesquisadores e simples interessados pela questão. Isto naturalmente envolve ainda, além de tradutores, terminológos, lexicógrafos, linguistas e literatos.

Mesmo que os passos para a automatização da geração de glossários bilíngues seja, no escopo deste trabalho, teórica e expositiva, acredita-se que a proposta de oferecer uma ordem sequencial gerará subsídios para a sua reprodução concreta, viabilizando seu domínio àqueles que estiverem interessados na recuperação de termos em corpora paralelos. A validação dos termos constituirá uma etapa científico-pessoal, que poderá eventualmente ser revista por especialistas do domínio estudado, examinada por peritos, ou mesmo ser dominada pelo tradutor a fim de garantir sua legitimidade. Isto marca os limites da máquina e a importância do insubstituível trabalho cognitivo: discernimento, intelectualidade, tecnologia, saber e experiências.

1.4 PERGUNTAS DE PESQUISA

Ao lançar a presente proposta de trabalho, algumas questões emergiram relativamente à proposta metodológica para extração terminológica, tais como:

(a) O processo de busca por correspondentes tradutórios, posto em operação através do uso de corpora poderia ser alcançado através da utilização das tecnologias que se têm atualmente disponíveis, ou seria necessário acrescentar programas que pudessem gerar dados de naturezas diferentes daqueles comumente obtidos?

(b) Qual a razão para que estudos já realizados se baseiem nas mesmas etapas metodológicas, porém não debatam a questão de busca por correspondentes tradutórios de forma automatizada? Tal tarefa seria realizável? Caso positivo, qual o grau de precisão e de confiabilidade dos resultados relativamente às necessidades do tradutor?

(c) Resultados satisfatórios podem ser alcançados por meio dos programas computacionais, no caso dessa pesquisa a ferramenta *GIZA++* em conjunto com o *WordSmith Tools*, fornecer auxílio ao tradutor em sua necessidade de terminologia *ad hoc*?

Naturalmente, algumas das respostas só serão alcançadas por meio de consulta a especialistas da área da medicina, tendo em vista o necessário conhecimento aprofundado da área para que se possa referendar correspondentes e equivalentes fornecidos pelo sistema de processamento.

Torna-se imprescindível evidenciar ainda que vários estudos similares já foram realizados em âmbito nacional, destacando-se Tagnin, (2003), Fonseca (2004), Fromm (2004), Teixeira (2005, 2010), Matuda (2010), outrossim, a indagação principal em propor uma metodologia de compilação de glossários, também não é nova, podendo ser encontrada nos trabalhos de Tagnin (2007), Tagnin e Fromm (2010), Santos (2010), e que têm seus pontos fortes na questão da busca de correspondentes tradutórios de forma manual; no entanto percebeu-se uma necessidade de colmatar a escassez desses estudos, já que essa etapa do processo pode, justamente, ser realizada de forma automatizada em virtude da urgência dos trabalhos em tradução, coincidindo com o advento das redes de informação em escala global. Não é coincidência que grandes *thesaurus* lexicais estejam disponíveis em rede e abertos ao público.

Observou-se que a maioria dos trabalhos como os de Fromm (2004), Fonseca (2004) Tagnin, (2003, 2007), Teixeira (2005, 2010), Tagnin e Fromm (2010), Matuda (2010), Santos (2010) focam suas ações na criação de glossários e bancos terminológicos e apresentam

frequentemente uma abordagem centrada em uma única direção, nas relações entre as línguas envolvidas, para a geração dos respectivos glossários – e neste caso específico, trata-se do inglês para o português. Geralmente visam à obtenção de listas de palavras-chaves nas quais o termo considerado fonte é geralmente aquele situado no texto de partida e, frequentemente, utiliza a língua inglesa que, como citada anteriormente, é amplamente tratada como língua franca há quase meio século.

Tendo ciência de outros estudos realizados, tal como apontado acima, e percebendo algumas das lacunas ainda presentes no trabalho daqueles autores, surge então o questionamento que nos leva a delimitar de forma mais específica o foco desta investigação, a saber: como propor uma metodologia para criação de glossários e bancos terminológicos, de modo que seja possível a geração de correspondentes tradutórios em ambas as direções relativamente às línguas envolvidas nos corpora comparados? Entende-se aqui por *direções* as relações do tipo: português/inglês, inglês/português. Nesse caso os percursos não seriam unidirecionais (i.e., $\Rightarrow$), mas sim bidirecionais (i.e.,$\Leftrightarrow$ biunívocos).

Ao se pensar em como realizar essa tarefa, sem a necessidade de executar buscas por possíveis traduções aos candidatos a termos no corpus de forma manual e com base nos corpora de estudo, pesquisando em dicionários bilíngues, ou até mesmo em corpora de referência, um obstáculo maior instaurou-se e, desde então, um novo questionamento emergiu, qual seja:

> *– Como propor uma metodologia automatizada para criação de glossários e bancos terminológicos focada em ambas a direções relativamente às línguas envolvidas nos corpora justapostos?*

Juntamente esse novo questionamento passou a ser um agente incoativo para uma nova noção sobre este estudo. Neste sentido, o desafio passou a ser a integração de ferramentas, àquelas que têm por finalidade o processamento da linguagem natural, já existentes para auxiliar no processo de extração terminológica, focado na busca automatizada por correspondentes tradutórios, formando uma proposta sequencial para tal tarefa. Percebe-se que ao longo do processo de reflexão os questionamentos foram se tornando mais específicos e, por conseguinte, contribuindo de forma decisiva para a delimitação das

fronteiras em cujo escopo da pesquisa se encerra, conduzindo o pesquisador a centrar seus esforços no problema principal, a saber: a automatização do processo de extração bilíngue.

Sabendo que as ferramentas oferecidas pela linguística de corpus foram desenvolvidas com o propósito de tratar da língua e proporcionar o desenvolvimento de modelos, com vistas a auxiliar o trabalho do tradutor, bem como do próprio compositor dos textos científicos e também dos estudiosos da língua, muitas dessas ferramentas por si só, quando não encaixadas em um processo sequencial, não oferecem aos usuários processos totalmente positivos em sua totalidade, sendo utilizados apenas para situações específicas e com capacidades reduzidas. Ao realizar a integração destas ferramentas, acredita-se que o tradutor/pesquisador contará com o trabalho de forma sequencial, sendo oferecidos todos os processos necessários à extração terminológica bilíngue. Além disso, ao ter acesso ao modelo adotado, poderá ele mesmo (o tradutor) propor melhorias na sequência das tarefas visando aperfeiçoar a obtenção dos resultados. Naturalmente, tal projeção poderia ocorrer em situação controlada, na qual se supõe que o utilizador e o *conceptor* dos programas pudessem dialogar a respeito da manipulação dos instrumentos para a obtenção de melhores performances.

Com relação à necessidade de desenvolvimento e aperfeiçoamento de ferramentas automatizadas para extração terminológica, visando integrar, aperfeiçoar e criar novos programas, seria possível questionar:

> *– Afinal, como realizar tal tarefa? Somente apresentando uma proposta de sequência linear de uso? Ou desenvolvendo um pacote completo de produtos para a extração terminológica bilíngue em todas as suas fases?*

Com os avanços da Linguística de Corpus enquanto campo de atividade e pesquisa, muitas ferramentas são – constante e progressivamente – desenvolvidas para o processamento da língua natural de forma cada vez mais ágil. No entanto, ferramentas quando percebidas de modo separado, isto é, isolado, podem gerar alguns equívocos. Supõe-se que muitos tradutores e pesquisadores, podendo extrair recursos empregando uma ferramenta, talvez ainda não suponham, ou não cogitem, a respeito da possibilidade de adoção de um grupo de programas em diálogo uns com os outros.

No caso do presente estudo, como já apontado, trata-se de uma proposta sequencial para o uso linear de programas com vistas à extração terminológica bilíngue, à geração de listas de palavras e listas de palavras-chave, ao alinhamento ao nível de sentença e palavra e à criação de mecanismos para a recuperação e análise de dados gerados através destes processos. Busca-se oferecer ao tradutor/pesquisador uma proposta monolítica, exatamente por preencher lacunas outrora presentes até mesmo em produtos, comercializados legalmente e visando às mesmas finalidades explicitadas acima.

Inicialmente, espera-se responder ao principal questionamento apresentado e, acredita-se, por consequência, vir a integrar ferramentas de processamento de corpora e perceber os cálculos de probabilidade empregados, de forma que possam ser elucidados – e elucidantes – durante o desenvolvimento desta investigação. Finalmente, que, por meio da integração de ferramentas adicionais, se torne possível alcançar o desenvolvimento de um sistema completo de extração terminológica, passível de ser utilizado de modo bidirecional relativamente àquelas que compõem os textos comparados.

1.5 OBJETIVOS DA PESQUISA

Como se observa, os objetivos desta investigação, condensados em seu título, se ramificam neste ponto do trabalho, derivando para realizações afins, indispensáveis na concretização da presente proposta. Assim, nesta seção, reafirma-se o objetivo geral e seus prolongamentos, tais como serão desenvolvidos ao longo deste trabalho. A seção divide-se em duas partes, a saber: (i) objetivo geral, e (ii) objetivos específicos, tal como segue abaixo.

1.5.1 Objetivo Geral

Visa-se propor uma ordem sequencial de aplicação de programas computacionais para tratar da questão da extração terminológica semiautomática por meio de ferramentas de análise estatística em corpora paralelos – inglês/português – destinados à geração de glossários bilíngues, com vistas à tradução de textos das ciências médicas.

1.5.2 Objetivos Específicos

a) Desenvolver uma proposta para a extração de glossários bilíngues a partir de corpora paralelos, visando preencher supostos *gaps* (lacunas) passíveis de aperfeiçoamento, com vistas ao processamento e à extração terminológica para a geração de glossários bilíngues;

b) elaborar programas computacionais específicos para a proposta de ordem sequencial como, por exemplo, gerador de listas de palavras, listas de palavras chaves por meio das abordagens de qui-quadrado, correção do qui-quadrado e da razão de verossimilhança, de modo a atingir resultados pretendidamente mais precisos;

c) analisar qualitativamente parcela dos resultados obtidos em relação ao par de línguas trabalhado com a finalidade de validar os termos recuperados de forma automática pela máquina.

No que concerne aos processos semiautomáticos, destaca-se a importância da preparação dos corpora tendo como ponto central o procedimento de alinhamento ao nível de sentenças, de forma a garantir maior confiabilidade na geração do alinhamento em termos de palavras.

Neste ponto, a intervenção e manipulação humana se fazem necessárias, posto que ainda não existe a possibilidade de reconhecimento de entidades dotadas de significação, tampouco de seus conteúdos específicos; ou seja, a máquina é totalmente capaz de identificar sequências de caracteres (palavras), mas ainda não possui habilidade para trabalhar sobre o léxico enquanto unidade de significação, tampouco sobre os limites em que se encerram proposições lógicas, porções significativas ou não.

Enquanto procedimentos automáticos, trata-se da aplicação de um conjunto de ferramentas integradas para a geração de (a) listas de palavras; (b) listas de palavras-chave; (c) listas de palavras alinhadas, através do alinhamento dito *ao nível de palavra*, cuja aplicação demanda a criação de programa específico para a recuperação dos alinhamentos produzidos com vistas à geração dos glossários bilíngues inglês/português.

Enquanto produto, trata-se de gerar e um glossário terminológico especializado, relativo à área das ciências médicas, baseado sobre textos da área, por meio do emprego da proposta

sequencial voltada exclusivamente à extração terminológica bilíngue em corpora paralelos – inglês/português – com vistas a contribuir, de forma indireta, na composição de materiais para a tradução de texto das ciências médicas.

Do ponto de vista técnico, busca-se trazer aportes para o aperfeiçoamento de ferramentas destinadas ao gerenciamento de glossários probabilísticos gerados pelo processamento estatístico dos corpora paralelos. Tais contribuições se integrariam, ainda que em caráter teórico, a Programas de Auxilio à Tradução (PATs), empregando formatos de arquivos de glossários compatíveis com os principais sistemas presentes no mercado, citando aqui particularmente o *WORDFAST*, *OMEGAT+* e o *TRADOS*.

Naturalmente, para fechar de forma menos árida a presente proposta, como já explicitado no terceiro objetivo específico (cf. "c"), destaca-se e examinam-se algumas unidades lexicais e/ou expressões mais salientes nos corpora investigados, de modo a discutir as relações estabelecidas pela máquina, comparativamente às análises realizadas por meio do raciocínio e julgamento humanos. Tenciona-se verificar se os resultados condizem com as projeções lançadas aqui em teoria, pois se acredita que as tendências atuais na área dos Estudos da Tradução é sublinhar o caráter qualitativo em paralelamente àquilo que se possa propor quantitativamente.

1.6 CONTRIBUIÇÃO E RELEVÂNCIA DO ESTUDO

Por apresentar um modelo de extração terminológica focado em dados de natureza linguística, a proposta visa gerar dados que, eventualmente, tanto possam interessar a pesquisadores quanto a tradutores que vislumbrem o uso de bases terminológicas como recurso adicional a suas atividades. Do mesmo modo, expor um *saber-fazer* àqueles que se interessa em conhecer o que subjaz ao processamento e que conduz aos equivalentes tradutórios realizados de modo automático. Mais gravemente, pensa-se nos pesquisadores que se iniciam na arte do processamento automático de línguas como conhecimento adicional às atividades de tradução de textos científicos.

Dentre os trabalhos realizados na área dos estudos da tradução, principalmente no Programa de Pós-Graduação em Estudos da Tradução (PPGET) da Universidade Federal de Santa Catarina, trata-se de mais uma inciativa com vistas a alavancar as discussões relacionadas como o

desenvolvimento de glossários, bancos terminológicos e dicionários bilíngues. Ao buscar automatizar processos e, ao mesmo tempo, atrair a atenção para resultados de natureza lexical, deriva-se à preponderância do analítico sobre o sintético, caso contrário se contemplaria procedimentos que pouco tem a ver com o *discurso*, base primária na composição dos gêneros textuais e, no caso em questão, dos traços que definem as línguas de especialidade.

Dentre os principais pontos a serem destacados, enquanto eventuais interesses da pesquisa aos profissionais da tradução, citam-se:

a) descrição integral da proposta sequencial para a extração terminológica bilíngue, focada no par linguístico inglês/português realizada sobre dados considerados de *larga escala*. Na verdade, um sistema integrado passível de ser reproduzido e aplicado sobre outros pares de idiomas e em outros campos do saber;

b) demonstrações de como as ferramentas já existentes no campo da linguística de corpus, como o *WordSmith Tools* e o *GIZA++* podem ser integradas, aperfeiçoando seu uso e, por extensão, valorizando seu uso por tradutores e pesquisadores;

c) apresentação de extração terminológica e geração de glossários por meio de inserções por modelo semiautomatizado podendo manter a precisão dos trabalhos manuais;

d) inclusão e apresentação das últimas ferramentas colocadas no mercado para a realização de processamento de línguas, especificamente destinadas à extração terminológica.

2 FUNDAMENTAÇÃO TEÓRICA

Tendo em vista o caráter multidisciplinar, inerentemente instalado na em relação à Linguística de Corpus, não se poderia dispor de um único referencial teórico teórica para a realização da proposta. Assim, procura-se contemplar as quatro correntes de estudo aqui contempladas e cuja integração parece incontornável. São elas: Estudos da Tradução e Corpora (ETC), Linguística de Corpus (LC), Terminologia (TER) e Extração Terminológica (ET). Na sequência, cada uma dessas linhas teóricas será brevemente discutida.

2.1 CONTEXTUALIZAÇÃO DA PESQUISA

O campo disciplinar – e também inter e multidisciplinar dos Estudos da Tradução, desde sua instituição como área do saber e fórum acadêmico de investigação, tem constituído um território ao qual muitos pesquisadores lançam seus esforços com o intuito de estabelecer modelos para pesquisa avançadas e para novos assentamentos de práticas científicas. Estudiosos como Holmes (1972, 1978, 1988), Lambert e Van Gorp (1985) e Hatim (2001) são nomes que exemplificam tal asserção e que se tornaram precursores deste campo disciplinar.

De acordo com Beeby (2000), de modo geral, ao iniciar um estudo sobre determinado tema, parece ser, inicialmente, essencial, "identificar o objeto a ser investigado e a razão pela qual se o está estudando". No caso em questão, a pesquisa insere-se no do campo disciplinar dos Estudos da Tradução e faz parte da ramificação das linhas de estudo propostas por Holmes (1972, 1988), situando-se exatamente na rubrica intitulada: *Programas de Auxílio ao Tradutor (PATs),* tal como pode ser observado na figura abaixo, na qual se representa os possíveis tópicos de discussão da, então, *nova* disciplina. Ao citar *Teorias da tradução restritas ao meio*, o autor destaca que estas ainda podem ser subdivididas entre as teorias da tradução realizadas por humanos, referindo-se aqui à tradução humana; as teorias da tradução realizadas por computadores, citando-se aqui a tradução automática; e as teorias da tradução, que estudam a junção entre a tradução humana e a tradução automática.

Ainda, circunscritos nos apontamentos de Holmes (1972/1978/1988, p. 181, 182), o autor realça que os diálogos e aproximações entre a tradução humana e a tradução automática pode ser interpretada como *a tecnologia servindo de auxílio ao tradutor humano* não somente na prática tradutológica efetiva, mas igualmente no âmbito das atividades de pesquisa com vistas aos Estudos da Tradução. Neste caso, as ferramentas de auxílio podem ser apresentadas sob as mais variadas formas, porém destacam-se duas rubricas, a saber: (a) as ferramentas de auxílio lexicográfico e de suporte terminológico e; (b) as gramáticas.

Inserindo a presente pesquisa nos postulados método-epistemológicos que Holmes (1972/1978/1988) colocou em pauta, desenvolve-se esta proposta ousando-se buscar promover ferramentas de auxílio ao tradutor, baseados sobretudo em aspectos terminológicos. Tal meta naturalmente exige postura específica, pontualmente delineada a responder o(s) objetivo(s) fixado(s), tal como se poderá observar no desenvolvimento desta proposta de estudo.

Destacamos ainda que ao citar as ferramentas de auxílio ao tradutor, acredita-se que Holmes (1972/1978/1988) provavelmente fazia referência às ferramentas disponíveis na época de apresentação de seu artigo, que apesar de seminal, reveste-se de certo caráter anacrônico. Tal descompasso temporal, também afeta a atualidade científica. Todavia, os progressos na área são compreensivelmente naturais diante da inexorabilidade dos avanços tecnológicos em um campo em que a velocidade das transformações dos produtos e processos ocorre de forma surpreendente e imprevisível. Atualmente as máquinas são passíveis de se tornarem obsoletas em 2 ou 3 anos, exigindo que se faça investimentos em novos produtos.

Os avanços tecnológicos, ligados à microinformática parecem ultrapassar, em certos campos, as adaptações humanas, pois não aguardam a preparação das gerações à recepção dos novos instrumentos. Por exemplo, um indivíduo que tenha nascido nos anos 1950, 1960, 1970, 1980 experimenta, hoje, um bombardeio de metamorfoses sem necessariamente se adaptar a todas elas. Paralelamente, sabe-se que atualmente as versões digitalizadas competem com livros em papel até mesmo nas bibliotecas, entretanto, ainda há compradores para as edições em papel e pessoas que, provavelmente, juram que tal prática nunca irá mudar.

Recentemente, com o anúncio de que a *Encyclopædia Britannica* cessaria sua coleção de volumes impressos após 244 anos[4], percebemos que os meios de divulgação de informação e pesquisa como enciclopédias, dicionários e, no que tange essa pesquisa, os tradutores on-line condenam às prateleiras materiais que outrora eram objetos de consulta a cada instante. Nas próprias palavras do presidente da companhia, Jorge Cauz, "Uma enciclopédia impressa se torna obsoleta no minuto em que é impressa. [...] Enquanto nossa versão online é atualizada constantemente[5]". Logo, as ferramentas de auxílio ao tradutor, às quais Holmes (Ibid.) fez referência, são naturalmente diferentes daquelas que projetamos, na atualidade, ao ler seus textos.

Os produtos atuais já não se baseiam mais em pequenos glossários ou dicionários, tampouco em simples processadores de textos com vistas à tradução, menos ainda em ferramentas de gerenciamento terminológico. Do mesmo modo, a própria noção de computador, de memória, de processamento, evolui a cada instante.

Os sistemas informáticos de suporte ao tradutor são – atual e prioritariamente – interativos e permitem não somente a navegação, mas aprimoramentos *à la carte*, gerando, por conseguinte, aparatos personalizados ao gosto dos utilizadores. Os efeitos idiossincráticos negativos, eventualmente gerados por ergonomias mal projetadas, podem doravante ser facilmente minimizados ou retificados. O próprio utilizador poderá interferir em suas fontes de consulta, tornando-as mais adaptadas aos seus usos. Seus dados pessoais poderão se tornar mais conviviais e mais ricos face às suas necessidades locais e imediatas. Para isso, se faz importante vislumbrar produtos em consonância com tais tendências, ou seja, sistemas abertos e *manipuláveis* em sentido benéfico e aperfeiçoador de ações, desenhado com extensões pontuais no sentido de promover a memória anexa, extensiva e artificial, para além das capacidades humanas ditas *naturais*. Eis o princípio básico da máquina: servir ao homem e se encarregar da mensagem (cf. Mc Luhan, 1969)

2.2 MAPEAMENTO DA PESQUISA

[4] http://www.telegraph.co.uk/culture/books/9142412/Encyclopaedia-Britannica-stops-printing-after-more-than-200-years.html

[5] A printed encyclopedia is obsolete the minute that you print it," Mr Cauz said. "Whereas our online edition is updated continuously."

Ainda com o intuito de propor um mapa para situar as pesquisas na área dos Estudos da Tradução, a obra *The Map – A Beginner's Guide to Doing Research* de Williams e Chesterman (2002) delineia as doze áreas de pesquisa contidas neste campo disciplinar. Dentre as doze apresentadas pelos autores, explicitadas na figura abaixo, a que nos interessa, particularmente, é aquela através de quais aspectos textuais poderão ser tratados de forma automática, por meio de programas computacionais desenvolvidos especificamente para tais fins.

Figura 1: The Map - A Beginner's Guide to Doing Research (WILLIAMS e CHESTERMAN, 2002)

Passou a ser tomado como *lugar comum* afirmar que as novas tecnologias da informática e da informação se tornam, cada vez mais, via incontornável para o tradutor moderno. Todavia, tal asserção ainda remete a panoramas plurais e um tanto opacos, pois, em geral, os cursos formais de capacitação para manuseio e uso destes novos recursos ainda são raros e destinados a públicos restritos. Neste sentido, parece tornar-se pertinente expor os detalhes subjacentes que conduzem aos produtos finais desejados pelo utilizador que optar por tais ferramentas. A experiência, seja ela sensível ou científica, pode abrir novos horizontes,

tanto para o aperfeiçoamento do uso, quanto para a modelação *à la carte* dos recursos de apoio ao tradutor.

2.3 ESTUDOS DA TRADUÇÃO E CORPORA (ETC)

Inicialmente, focamos dois pontos fundamentais em relação aos Estudos da Tradução enquanto disciplina. Primeiramente, podemos observar que com o advento das tecnologias foi possível melhorar o acesso às informações e, consequentemente, a rapidez no processamento das línguas naturais. A aproximação entre povos e culturas exigiu a criação de formas de comunicação mais ágeis e integradas. Naturalmente, do ponto de vista político-crítico, tais aportes contemplam primordialmente as línguas majoritárias, o que permitiu à língua inglesa firmar-se ainda mais como uma das principais línguas de comunicação da atualidade. Logo, tendo em vista sua adoção ampla no universo político e econômico, observa-se a necessidade de integração dos mais diversos idiomas para formar par com o *inglês*, aumentando a demanda por traduções envolvendo os mais variados pares linguísticos. Atualmente, entendemos que a tradução apresenta um papel crucial nas relações de todas as ordens: acadêmicas, informativas, diplomáticas, culturais.

Nesse sentido, embora se trate de *lugar comum*, parece possível afirmar que o advento da globalização como processo político, mesmo que criticável em diversos pontos que fogem ao escopo dessa proposta, tem sido um dos principais responsáveis pela explosão das demandas prementes no campo da tradução, só igualável à instauração da imprensa durante o Renascimento. Essas causas e suas decorrentes consequências políticas, sociais e econômicas, em certo sentido, podem ser aceitas como influenciadoras da instauração da disciplina de Estudos da Tradução. Os motivos literários, discursivos, científicos e culturais naturalmente constituem a base primeira a ponto de existir uma História da Tradução. Todavia, trata-se aqui de pautar a língua como instrumento de divulgação do saber em suas mais diversas vertentes.

Em síntese, os Estudos da Tradução constituem uma disciplina acadêmico-científica que se volta aos estudos teóricos, críticos e práticos relativamente aos diversos fenômenos que permeiam a interpretação e a tradução. Implicitamente, sua natureza está ligada ao multilinguismo e à inter e multidisciplinaridade, contemplando as mais diferentes expressões decorrentes da linguagem como manifestação. Atualmente, a disciplina ultrapassa as concepções de Jakobson

(1959/2000), que estratifica a tradução como intralinguística, interlinguística e intersemiótica. As fronteiras estanques progressivamente cedem lugar aos estudos do discurso que, por sua vez, circunscreve em seus domínios as expressões estéticas, as mais variadas.

Se a tradução e os estudos literários continuarão a constituir a base da história da tradução, as novas tecnologias doravante abrem vias para a consideração dos eventuais aportes que podem trazer para que se tenham novas e diferentes visões a respeito da noção de texto. Muitos avanços nesse campo já se tornaram consubstanciais ao ato de ler e de escrever. Não se atenta mais para as vantagens da digitação virtual em comparação com a datilografia. Imaginemos, para não ir tão longe, redigir um trabalho com páginas numeradas em uma máquina de datilografia, tendo que suprimir uma delas. Isto exigiria a repaginação do conjunto e, como consequência, muitos dias de trabalho. Do mesmo modo, o trabalho do tradutor com suportes informáticos, cada vez mais se torna prática corrente e progressivamente consubstancial a seu exercício.

O campo disciplinar dos estudos da tradução passou a constar dos currículos com a instauração da disciplina, ocorrida em meados do século XX, quando Holmes apresentou um artigo intitulado "*The Name and Nature of Translation Studies*" no Setor de Tradução do Terceiro Congresso Internacional de Linguística Aplicada, realizado em Copenhagen em 1972. No referido artigo, Holmes descreve os campos que os estudos da tradução cobririam, dividindo-os em Puro e Aplicado. Contudo, para fins desta pesquisa, seguindo o mapeamento de Toury (1995), baseado no artigo de Holmes (1972, 1988), foi utilizada a teoria Aplicada, Ferramentas de Auxílio à Tradução (ver Figura 2).

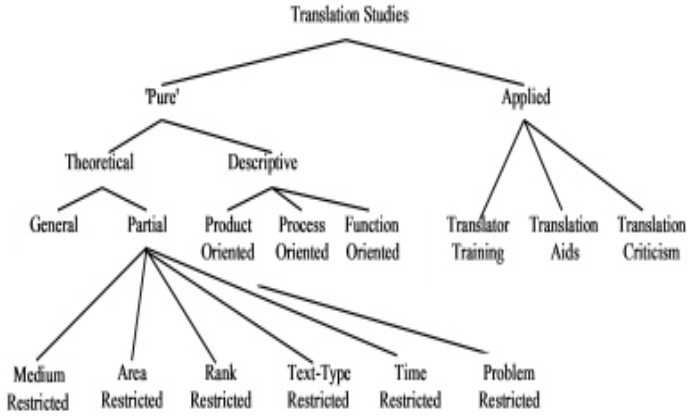

Figura 2: Mapa dos Estudos da Tradução (baseado em HOLMES, 1972)

Recentemente, com os avanços tecnológicos mais modernos, acredita-se que se passou a estudar as traduções em grande parte por meio de textos digitalizados, sobretudo no que concerne aos textos científicos. Neste sentido, sabe-se que os corpora, no caso presente: textos em formato eletrônico, relativos a domínios específicos, apresentam importante impacto nos estudos da tradução científica. Essa abordagem parece ter possibilitado ao tradutor o armazenamento de enormes quantidades de dados que conduzem à investigação, análise e comparação da natureza dos elementos neles contidos, entre outros fenômenos que fornecem percepções nos processos tradutórios, bem como para as análises linguísticas que lhes são inerentes.

Além disso, o fato de os corpora apresentarem contextos ditos *reais* (ou autênticos) representa um dos importantes recursos para as pesquisa em tradução. O esforço em se penetrar mais na natureza do texto e de seus contextos traz possibilidades de escolhas ao tradutor, auxiliando-o em seus processos conscientes de escolha e seleção, bem como em sua posição inconsciente ao ativar informações latentes que precisam emergir por meio de insumos, agindo como desencadeadores de tomada de decisões e de estratégias tradutórias. Como já observado, não acreditamos que a Linguística de Corpus possa resolver problemas envolvendo ambiguidades e escolhas, mas pode constituir uma via capaz de minimizar, por exemplo, tarefas de busca, de definição de hierarquias, apontando sobretudo para denotações, conotações, associações e exemplos. Tais índices, embora sejam contemplados nos

dicionários analógicos (em papel), além de precisarem ser manuseados materialmente, em geral, se não forem dedicados à línguas de especialidade, poderão conter muitos *gaps* e generalizações.

Cabe, humildemente, destacar a configuração formativa em tecnologia e terminologia deste pesquisador, ao ter sido contemplado com a oportunidade de desenvolver estudos concomitantemente em Terminologia, Linguística de Corpus e Estudos da Tradução Baseados em Corpora. Cite-se, a título ilustrativo, os trabalhos de Bartholamei e Vasconcellos (2008), também apontado por Fromm (2007).

Faz-se igualmente importante remeter o leitor ao estudo de Gonçalves e Machado (2006) a respeito dos conhecimentos necessários para a formação de um tradutor. Os autores apresentam um panorama do ensino da tradução no Brasil, discutindo a questão da competência tradutória. Na Figura 3, de autoria dos autores em questão, é demonstrada a importância de duas áreas fortemente focadas nessa investigação, o item de número 8 (oito) *Terminologia* e o item de número 12 (doze) *Tecnologias que podem ser aplicadas à tradução*.

Figura 3 Competências Tradutórias em Cursos de Tradução no Brasil (GONÇALVES; MACHADO, 2006)

Ambas as áreas destacadas nesta investigação, presentes no estudo de Gonçalves e Machado (2006), demonstram a importância em se disponibilizar ferramentas complementares àquelas oferecidas pelas disciplinas tradicionais agregando à evolução e ao desenvolvimento das novas tecnologias e aplicadas a disciplinas como a Terminologia, uma vez que a Linguística e os Estudos da Tradução já estariam sendo contemplados com essa evolução.

Os Estudos da Tradução Baseados em Corpora tiveram início com os trabalhos de Baker (1995) em seu artigo intitulado: "*Corpora in translation studies: An overview and some suggestions for future research*", considerado seminal pela maioria dos pesquisadores, teóricos

e estudiosos da área. Para Baker (Ibid.), a possibilidade que os pesquisadores teriam de investigar a "natureza dos textos traduzidos como eventos comunicativos mediados" foi lançada como argumento de base para o uso de corpora nos estudos da tradução. Considerando que as ferramentas da linguística computacional têm evoluído consideravelmente, isto faz com que a análise de grande volume de material textual se torne não somente possível, mas também que não encontre limites. Referimo-nos particularmente à quantidade de fenômenos e elementos emergentes, situados no escopo de trabalhos como o presente.

A proposta inicial de Baker (1993, 1995, 2000) sobre a utilização de corpora nos estudos da tradução sugere uma diferenciação na tipologia do corpus usado para pesquisas relacionadas à tradução. A autora destaca a importância da tipologização quanto aos modelos de corpus que podem ser utilizados de acordo com os objetivos específicos da pesquisa. Para ela, o uso de corpora nos estudos da tradução é dividido em três categorias, são elas:

(i) corpus paralelo;

(ii) corpus comparável;

(iii) corpus multilíngue.

Para a definição sobre a tipologia de corpus utilizada neste estudo, remetemos o leitor à seção Tipo de Corpus.

Embora se faça uso de diversas ferramentas da linguística de corpus, Baker (Ibid.) destaca a importância do desenvolvimento de ferramentas específicas para a investigação de fenômenos ocorrentes da tradução, na relação entre texto e texto traduzido, como somente aqueles percebidos em textos traduzidos. Além disso, Tymoczko (1998, p. 97) sugere a adaptação das novas tecnologias para o uso no novo campo de pesquisa dedicado aos Estudos da Tradução.

2.4 LINGUÍSTICA DE CORPUS (LC)

A Linguística de Corpus - ora utilizada como teoria, ora como metodologia - refere-se a uma disciplina cujo objetivo é a realização de análises linguísticas a partir de grandes volumes textuais. Segundo Berber Sardinha (2004), a Linguística de Corpus preocupa-se com a

coleta e exploração de corpora selecionados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística; como tal, explora esses fenômenos linguísticos por meio de evidências empíricas extraídas por computador.

Para a presente proposta de trabalho, vislumbramos a Linguística de Corpus (LC) como abordagem metodológica, posto que expõe métodos que possibilitam a manipulação e o processamento do material textual, oferecendo ferramentas necessárias para que o pesquisador seja capaz de examinar minuciosamente os dados de seu material textual.

Nesse sentido, observa-se que o computador assumiu um caráter essencial quando adotado como ferramenta de auxílio à pesquisa. Na área da Linguística de Corpus, seu papel foi de tal importância que se fez necessário à criação, nos meados dos anos 1970, de uma nova disciplina, denominada de Linguística Computacional, que posteriormente se ramificou tanto em razão de demandas específicas quanto em função do caráter generalizante do termo inicial.

O desenvolvimento de ferramentas que possibilitassem análises detalhadas de aspectos linguísticos em um texto foi um dos principais fatores para o crescimento e a disseminação dessa dita *metodologia*. Tamanha foi sua aceitação que as ferramentas e métodos criados nesse contexto serviram também como base para criação de outras disciplinas, como por exemplo, Estudos da Tradução Baseados em Corpora.

De acordo com Leech (1992), a Linguística de Corpus se estabelece como um meio essencial para uma nova forma de busca por informações e, também, como portal para novos modos de pensar a respeito das línguas e das linguagens. Despontando como eficiente fórum de discussões para investigar fatos envolvendo estas duas entidades essencialmente abstratas, possibilita o manuseio de milhões de dados de maneira monolítica e, concomitantemente, destacando unidades, posto que se centra essencialmente sobre as formas sem, no entanto, desconsiderar significações e sentidos (cf. ocorrências, colocações, etc.). Além de acelerar o manuseio de grandes quantidades de informações, ainda pôs a matemática a serviço da linguística, mostrando ser possível gerar modelos para o tratamento de comportamentos linguísticos, tal como a frequência e os cálculos estatísticos como de qui-quadrado e de razão de verossimilhança. A mencionada multidisciplinaridade da Linguística de Corpus torna-se, acreditamos, inquestionável.

Contando com a utilização de programas computacionais, fenômenos antes dificilmente detectados passaram a ser explicitados,

tornando-se foco de investigações. Dentre os principais elementos levantados na utilização dos programas computacionais pode-se destacar a probabilidade. De acordo com Berber Sardinha (2004), a Linguística de Corpus trabalha circunscrita em panorama conceitual formado por abordagens empiristas e uma visão da linguagem como sistema probabilístico. A posição de Berber Sardinha (Ibid.) toma por base a proposta de *Halliday*[6] que, por sua vez, procura seguir uma filosofia empirista baseado em probabilidades e na observação dos dados. Já *Chomsky*[7] apresenta uma visão racionalista buscando a linguagem como possibilidade.

> A linguística chomiskyana gerativista enfatiza a determinação de quais agrupamentos sintáticos são possíveis (permissíveis) dado o conhecimento que um falante nativo possui da língua. Já a linguística hallidayana descreve a probabilidade dos sistemas linguísticos, dados os contextos em que os falantes os empregam. (BERBER SARDINHA, 2004, p. 30).

De acordo com a visão *hallidayana* existiria uma probabilidade de determinadas ocorrências acontecerem em dados contextos. Nesse sentido, para saber qual a probabilidade de um traço ou de uma estrutura aparecer, responde-se à abordagem empírica que evidencia determinados padrões apresentados pela língua. Esses cálculos realizados sobre os dados linguísticos, em suas probabilidades e evidências empíricas, possibilitam aos pesquisadores uma visão detalhada do comportamento de cada elemento linguístico destacado no corpus em estudo.

Para o cálculo e geração dos dados resultantes do processamento, ferramentas desenvolvidas *à la carte* auxiliariam no processo. Uma das ferramentas mais utilizadas no campo da Linguística de Corpus é o *WordSmith Tools* (Scott, 1996, 1997, 1999, 2004, 2008). Sobre ferramentas específicas de processamento do corpus, solicita-se ao leitor que consulte a seção Ferramentas de Processamento. Para entendimento dos cálculos de probabilidade, consulte a seção Cálculos de Probabilidade.

---

[6] http://pt.wikipedia.org/wiki/Michael_Halliday

[7] http://pt.wikipedia.org/wiki/Noam_Chomsky

Das considerações acima, pode-se concluir que as informações advindas da linguagem natural, registrada em uso, em situações reais, são fontes de investigação relevantes. Saber, por exemplo, qual a probabilidade de um traço ou estrutura linguística ocorrer em determinado contexto é um dado importante para o tradutor. A observação empírica e a obtenção de cálculos de frequência de ocorrência desses traços linguísticos, sejam referentes à forma, ao léxico (forma + sentido), sejam de ordem sintática, semântico-discursiva (no patamar qualitativo realizado pelo pesquisador), auxiliam na análise dos padrões estabelecidos no âmbito da língua de especialidade estudada.

A busca por evidências de que a língua tende a funcionar de forma padronizada, ou seja, de que existe certa regularidade nas associações entre as unidades linguísticas, consiste do principal motivo que atrai os estudos sobre a linguística de corpus, tal como se pode observar nesta citação de Hunston (2002):

> Os padrões de uma palavra podem ser definidos como todas as palavras e estruturas com as quais são regularmente associados e que contribuam para o seu significado. Um padrão pode ser identificado se uma combinação de palavras ocorre com frequência relativa, se é dependente de uma palavra específica, e se há um significado claro associado. (HUNSTON, 2002, p. 37, tradução nossa)[8].

Ainda, de acordo com Hunston (2000), o procedimento de investigações dos padrões de uma palavra é naturalmente analisado por meio de linhas de concordâncias que envolvem seleções aleatórias extraídas por meio de ferramentas eletrônicas, desenvolvidas para o processamento e análise dos dados. Como exemplo, citamos o *WordSmith Tools* (ver Seção Ferramentas de Processamento). Um padrão lexical, por exemplo, poderá ser identificado se uma combinação de palavras ocorrer de forma significativamente frequente.

[8] The patterns of a Word can be defined as all the words and structures which are regularly associated with the Word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it.

Sinclair (1991), Berber Sardinha (2004), Partington (1998) e Tagnin (2005), classificam essas padronizações em três categorias:

(i) colocação;

(ii) coligação, e

(iii) prosódia semântica.

A colocação corresponde à combinação lexical de elementos linguísticos de duas ou mais palavras que normalmente fazem "companhia" umas as outras dentro de um *pequeno* espaço no texto. Por exemplo, a palavra *causar* que normalmente coocorre com as palavras *problema, sofrimento, tristeza, prejuízo, danos, morte* etc. Estudar essas coocorrências nos fornece entendimentos mais aprofundados do significado de uma dada palavra em contexto e da forma mais comumente utilizada por falantes nativos. Por exemplo, Kennedy (1991, p. 107) estudou o uso da palavra *between* e *through*, cujo uso, segundo apontado em seu trabalho, não se distingue facilmente nos livros. Por meio do estudo dos colocados descobriu-se que *between* é geralmente utilizado depois de substantivos como *differences, distinction, agreement* e *meeting*, por sua vez, *through* seria mais frequente depois de verbos, tais como: *go, pass, run* e *fall*.

Além das coocorrências entre dois ou mais itens lexicais, ainda podemos observar coocorrências entre itens lexicais e gramaticais. Nesse caso, Tagnin (2005) chama essa associação de *coligações*. Mais precisamente, a coligação corresponderia à combinação de duas ou mais palavras em que o colocado seria um elemento gramatical. Alguns exemplos de coligações citados por Tagnin (Ibid.) são: em inglês: *look at, mad about*; em português: *obedecer a, cumpridor de.* Destacamos que a interpretação dessas coligações e colocações depende da carga semântica que acompanham os itens lexicais ou gramaticais. Estas são chamadas de prosódia semântica que, segundo Berber Sardinha (1999), pode apresentar sentido positivo, negativo ou neutro. A sua determinação depende tanto do que a palavra representa, quanto do contexto em que ela é utilizada. Por exemplo, referenciando aqui o estudo de Stubbs (1995), o verbo citado acima, *causar*, será negativo no contexto: *O acidente causou prejuízos ao dono do automóvel*. No entanto, positivo para: *O nascimento de Helena foi causa de alegria e festa na casa dos Souza*.

Destacamos a relevância dos padrões para a interpretação dos textos. As relações entre os padrões e as informações podem ser percebidas e, a partir delas, são possíveis, pois, em princípio, a experiência e o conhecimento não estão relacionados a palavras individuais, mas sim a combinações entre elementos linguísticos.

> Um corpus é uma coleção de textos, projetado com base em um objetivo, geralmente de ensino ou pesquisa. [...] A corpus não é algo que um falante faz ou sabe, mas algo construído por um pesquisador. É um registro de desempenho, geralmente de muitos usuários diferentes, e projetados para ser estudado, para que possamos fazer inferências sobre o uso da linguagem típica. Porque fornece métodos de observação padrões de um tipo que têm sido percebido pelos críticos literários, mas que não foram identificados empiricamente, o estudo assistido por computador de grande corpora talvez possa sugerir uma maneira de sair dos paradoxos do dualismo. (STUBBS, 2001, p. 239-240, tradução nossa)[9]

Como visto em Stubbs (Ibid.), relações entre os padrões e as informações podem ser percebidas por meio das colocações, coligações e prosódia semântica. As palavras mantêm relações que devem ser percebidas para que se possa extrair do texto informações relevantes à sua efetiva interpretação. Neste sentido, a investigação das palavras-chave de um determinado contexto torna-se necessária ao entendimento do próprio contexto. Na sequência, destaca-se outro aspecto estudado na Linguística de Corpus, a chavicidade.

Na Linguística de Corpus a chavicidade (*keyness*) é calculada por meio da frequência da palavra no contexto. Berber Sardinha (2004) argumenta que um item lexical pode ser identificado como palavra-

[9] A corpus is a collection of texts, designed for some purpose, usually teaching or research. […] A corpus is not something that a speaker does or knows, but something constructed by a researcher. It is a record of performance, usually of many different users, and designed to be studied, so that we can make inferences about typical language use. Because it provides methods of observing patterns of a type which have long been sensed by literary critics, but which have not been identified empirically, the computer-assisted study of large corpora can perhaps suggest a way out of the paradoxes of dualism. (Stubbs, 2002, p. 239-240)

chave em um *corpus de estudo* (ver seção Corpus de Estudo) se ocorrer com frequência considerável ao ser estatisticamente comparado com outro corpus, usado neste caso como *corpus de referência* (ver seção Corpus de Referência). Esses dados linguísticos são utilizados para extrair informações relevantes do texto, principalmente para levantar dados que deem embasamento necessário às escolhas de candidatos a termo na elaboração de glossários ou dicionários.

Entre os fatores estatísticos, podemos considerar que os dados analisados de forma quantitativa por meio de métodos empíricos, comparando-se com o método racionalista (intuitivo), apresentam pontos positivos, pois fornecem embasamentos concretos e confiáveis. As pesquisas baseadas ou guiadas por corpus nos permitem responder a muitas perguntas, como por exemplo:

a) quais as palavras mais frequentes em um determinado contexto;

b) quais as principais diferenças entre um texto falado e um texto escrito (dentro de uma mesma especificidade da língua);

c) quais os tempos verbais mais utilizados na fala e/ou na escrita;

d) quais preposições seguem determinados verbos;

e) quais as palavras mais utilizadas no inglês formal *versus* informal;

f) com qual frequência se utiliza expressões idiomáticas;

g) quantas palavras um aprendiz deve dominar para participar de uma conversa cotidiana ou quantas palavras um falante nativo utiliza em uma conversação cotidiana em dado contexto, etc.

Em síntese, a utilização do computador para o desenvolvimento dos estudos baseados ou guiados por corpus é de extrema importância, pois a máquina pode ter por tarefa buscar, recuperar, selecionar e calcular instantaneamente, oferecendo ao pesquisador a possibilidade de compreender e explicar fatos de maneira fundamentada e segura, uma vez que se referem a procedimentos básicos em termos de processamento de dados. Entre os exemplos de utilização dos corpora citados acima, ainda encontram-se diversos profissionais envolvidos que utilizam a metodologia de linguística de corpus em suas atividades, por

exemplo: lexicógrafos, terminólogos, professores, estudiosos e profissionais da tradução, linguistas, aprendizes de idiomas, legendistas etc.

2.4.1 Abordagens de Análise em Linguística de Corpus

A abordagem de análise de corpus pode ser realizada de duas maneiras, uma delas é conhecida como *abordagem baseada em corpus* (*corpus-based*). A outra é denominada como *abordagem dirigida ou guiada pelo corpus* (*corpus-driven*). A primeira, baseada em corpus, remete a uma metodologia que tem por finalidade expor, testar ou exemplificar teorias e ideias elaboradas antes de o corpus tornar-se disponível. Neste sentido, Tognini-Bonelli (2001) parte à análise do corpus de modo a descobrir usos, e buscar resultados para verificar seus pressupostos iniciais nos dados do corpus, ou seja, se esses dados são utilizados como meio de validação e quantificação da teoria e descrição linguística.

> Um corpus pode ser definido como uma coleção de textos considerados representativosa de uma determinada língua e agrupados de modo que podem ser usados para a análise linguística. Geralmente, a premissa é que a língua armazenada em um corpus é de ocorrência natural, que é coletada de acordo com critérios explícitos, com um propósito específico em mente, e com a pretensão de representar a maior parte da língua selecionada de acordo com uma tipologia específica. [...] Em geral, há um consenso que um corpus lida com a língua natural, autêntica. (TOGNINI-BONELLI, 2001, p. 2, tradução nossa)[10]

[10] A corpus can be defined as a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis. Usually the assumption is that the language stored in a corpus is naturally-occurring, that it is gathered according to explicit design criteria, with a specific purpose in mind, and with a claim to represent larger chunks of language selected according to a specific typology. […] in general there is consensus that a corpus deals with natural, authentic language.

Biber et al. (1998) defendem que o maior objetivo da abordagem baseada em corpus não é apenas relatar descobertas sobre o uso dos padrões da língua, mas também explorar a importância delas. Em contraste à abordagem baseada em corpus (*corpus-based*), a abordagem dirigida ou guiada pelo corpus (*corpus-driven*) configura-se de forma que os textos sirvam como base empírica a partir da qual os pesquisadores extraem dados e detectam fenômenos linguísticos sem expectativas prévias. Os problemas a serem detectados nesta abordagem surgirão, via de regra, durante a análise do corpus.

Segundo Tognini-Bonelli (2001), a observação das linhas de concordância, ou seja, dos padrões remete o pesquisador ao levantamento de hipóteses, a generalizações e, por fim, à unificação das descobertas em teorias, a qual reflete a evidência destacada no corpus. Logo, pode-se dizer que o processo metodológico da abordagem dirigida ou guiada pelo corpus (*corpus-driven*) segue uma sequência definida por quatro passos principais, quais sejam: a observação, a hipótese, a generalização e a construção da teoria que, consequentemente, é influenciada pela experiência e intuição do pesquisador.

2.4.2 Corpus *versus* Corpi *versus* Corpora

A diferenciação entre corpus e corpora por vezes pode parecer confusa, sobretudo em se considerando a semelhança formal concernente à nomenclatura. Similarmente, além da diferenciação entre Corpus e Corpora, uma nova nomenclatura é adicionada, Corpi. Para evitar eventuais confusões a respeito da diferenciação entre os termos utilizados nesse trabalho, emprega-se apenas corpus e corpora. O critério desta escolha será, pois, explanado e referenciado, para maior esclarecimento e fluidez, durante a leitura do texto.

Corpus é uma palavra derivada do latim. Em Latim, existem diferenciações para formas flexionadas para cada classe de palavras. Corpus é uma forma da terceira declinação, a qual se flexiona como pode ser observado no quadro abaixo:

Classe	Singular	Plural
Nominativo	corpus	Corpora
Genitivo	corporis	Corporum
Dativo	corpori	Corporibus
Acusativo	corpus	Corpora

Vocativo	corpus	Corpora
Ablativo	corpore	Corporibus

Quadro 1: Corpus – Declinações

Partindo da origem latina da palavra, adota-se o nominativo para a representação do primeiro termo. Sendo assim, nesse trabalho a utilização da nomenclatura para fazer referência às formas singular e plural da palavra *corpus* será *corpus* e *corpora*, respectivamente. Para aqueles pesquisadores e estudiosos, cuja referência recai sobre o termo *corpi* para indicar o plural de corpus, referenciem corpora como *corpi*. Como o termo universalmente utilizado: corpus é, como visto, originário do latim, a escolha também contemplará leitores de outros idiomas, e não somente do português.

2.4.3 Definição de Corpus/Corpora

Desde o início da utilização de corpus para estudo linguístico e tradutológico, diversos autores têm apresentado definições para essa terminologia amplamente empregada. À ótica de um dos principais precursores da utilização desta metodologia na área da linguística: Biber (1998), por exemplo, a definição de corpus remete a: uma grande coleção de textos baseada em princípios naturais, ou seja, um conjunto de material textual produzido de forma natural. Uma definição ampliada seria a seguinte:

> [...] um corpus é uma coleção de textos naturalmente ocorrentes na língua escolhida para caracterizar um estado ou variedade de uma língua. Na linguística computacional moderna, um corpus geralmente contém milhões de palavras: isso ocorre porque é reconhecido que a criatividade da linguagem natural leva a uma imensa variedade de expressão tal qual é difícil isolar os padrões recorrentes que são os traços para a estrutura lexical da língua. (SINCLAIR, 1991, p. 171, tradução nossa)

Como proposto por Sinclair (Ibid.), o corpus remete a um conjunto de textos produzidos de forma natural, que se representa

através de suas características linguísticas, isto é, de aspectos do idioma aos quais tais textos remetem. Na área dos estudos da tradução, encontra-se outra definição importante:

> (i) corpus significa, essencialmente, uma coleção de textos armazenados de forma que sejam legíveis aos computadores e que possam ser analisados automática ou semiautomaticamente em diversas maneiras, (ii) um corpus não se restringe mais a "escritos", pois inclui tanto falados como também escritos, e (iii) um corpus pode incluir uma grande quantidade de textos a extraídos de várias fontes, de muitos escritores e oradores e de uma infinidade de tópicos. (BAKER, 1995, p. 225, nossa tradução).

Como se percebe na definição proposta por Baker (Ibid.), o núcleo da explanação permanece similar a outras classificações, no entanto, restrições são adicionadas, tal como o armazenamento do material contido no corpus em formato eletrônico, codificado de forma que se torne possível sua leitura por sistemas computacionais. Ainda em Baker (Ibid.), o conjunto de textos constituintes do corpus não apenas é composto de material textual escrito, mas também de materiais falados, coletados, codificados e transcritos, de forma, por exemplo, a serem *lidos* por aplicativos baseados em computadores.

Em EAGLES (1996), a definição de corpus ainda pode ser entendida como porções da língua, não necessitando ser entendido como uma generalização da língua. Logo, um corpus pode ser tomado como uma coleção de partes da língua – selecionadas e ordenadas de acordo com critérios linguísticos explícitos – a fim de ser usado e aceito como amostra da língua.

No relatório do EAGLES (Ibid.) em um corpus computacional, o material nele contido precisa ser codificado sob alguns padrões definidos:

> [...] um corpus computacional é aquele codificado de forma padronizada e homogênea para [...] as tarefas de recuperação. Suas partes constituintes

> da linguagem são documentadas quanto à sua origem e procedência. (SINCLAIR, 1996)[11]

Para os fins específicos desta investigação, a definição de corpus se baseará na compilação daquelas apresentadas pelos diversos teóricos acima citados. Por conseguinte, corpus se define como um conjunto de materiais linguísticos, coletados de fontes variadas; falados, escritos, sinalizados, interpretados, entre outros; armazenados em formato eletrônico e codificados de acordo com padrões previamente estipulados; processados tanto automática quanto semiautomaticamente; com a finalidade de se extrair dados para análise de fenômenos linguísticos e tradutológicos.

Ao consultar um corpus é importante saber qual a finalidade da pesquisa. Atualmente existem diversos tipos de corpora para pesquisa. Eis alguns deles, tratados em maiores detalhes em seguida. A saber: corpus geral, especializado, aprendiz, pedagógico, histórico, monitor, contemporâneo, paralelo, comparável, corpus dinâmico, estático, de amostragem, corpus monolíngue, corpus bilíngue e multilíngue, corpus de estudo, e corpus de referência.

2.4.4 Corpus e Corpora Gerais

Atualmente, o tipo de corpus encontrado mais facilmente é o corpus de língua geral. De acordo com Sinclair (1991) são largamente volumosos e homogêneos, mas contêm variedades de fontes linguísticas. Essa diversidade de fontes constitui um fator importante, pois garante amostragens mais eficientes.

Os Corpora Gerais são compostos por textos que se pressupõem serem representativos da linguagem cotidiana, isto é, diária, e não especializada. Esse tipo de corpus também é normalmente utilizado como corpus de referência por ser vasto em conteúdo e relativamente representativo no que se refere à língua geral. Apesar de nenhum corpus representar todas as possibilidades existentes da língua, o corpus de língua geral fornece aos usuários vasta amostra de variações

[11] A computer corpus is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.

linguísticas. O *British National Corpus* (*BNC*) e o *American National Corpus* (*ANC*) são alguns exemplos de corpus de língua geral.

2.4.5 Corpus e Corpora Especializados

Ao contrário, corpora especializados contêm textos de uma dada especificidade (campo disciplinar específico como, por exemplo, medicina, direito, educação, etc.) e espera-se que sejam representativos de acordo com os propósitos da pesquisa. Ambos os corpora, Geral e Específico devem ser representativos, mas o corpus geral e especializado é observado de forma diferente, pois enquanto o corpus geral abrange variados gêneros, o corpus específico é fechado, focando apenas determinado assunto. Quanto mais representativo de determinada área de estudo, mais evidências linguísticas se revelarão ao pesquisador. Esses corpora são, geralmente, criados para responderem a perguntas bem específicas. Pode-se citar como exemplo de corpus especializado o material da área jurídica do Parlamento Europeu disponível na interface do *NATOOLS* (2003).

2.4.6 Corpus e Corpora de Aprendizes

Semelhantemente, os corpora de aprendizes são corpora especializados que contêm textos escritos ou transcrições da fala usada por aprendizes que estejam aprendendo e desenvolvendo habilidades para o domínio de nova(s) língua(s). Esses corpora são normalmente etiquetados e evidenciam os equívocos que os estudantes apresentam no processo de aprendizagem/aquisição. Segundo Granger (2003) um corpus de aprendizes bastante conhecido é o *International Corpus of Learner English* (ICLE), que contem textos escritos por aprendizes em 14 idiomas diferentes.

2.4.7 Corpus e Corpora Pedagógicos

Os corpora pedagógicos contêm amostras da língua utilizada no cenário de sala de aula, incluindo textos de livros, transcrições de interações de ambientes de ensino, qualquer texto escrito ou transcrição da fala realizada em ambiente de aprendizagem. Segundo Hunston (2002) os corpora pedagógicos são materiais compostos por todas as

modalidades de língua às quais o aprendiz é submetido ou exposto por meio de textos que sejam parte integrante do material didático-pedagógico. O objetivo deste tipo de corpus é verificar se ele se assemelha ao corpus autêntico e se é útil na comunicação. Constitui base para se examinar as dinâmicas que se estabelecem entre professor e estudante ou, ainda, como uma ferramenta autorreflexiva para o desenvolvimento das práticas docentes. Podemos citar o projeto *COBUILD* dirigido por John Sinclair (ver Sinclair, 1987), desenvolvido com o objetivo de produzir materiais mais realísticos para serem utilizados no ensino de línguas em sala de aula.

2.4.8 Corpus e Corpora Históricos e Contemporâneos

Dentre os corpora que se enquadram em outros tipos, citam-se os corpora históricos. Segundo Berber Sardinha (2002), tal material abrange e remete a fatos ocorridos no passado, contrariamente ao corpus contemporâneo que concerne à acontecimentos do presente.

2.4.9 Corpus e Corpora Monolíngues, Bilíngues e Multilíngues

Corpora monolíngues são aqueles que se constituem de materiais linguísticos dispostos em um único idioma. Os Corpora Bilíngues envolvem duas línguas e os Multilíngues são compostos por três ou mais línguas.

Na literatura encontram-se autores que preferem empregar o termo *corpus comparável* para fazer referência ao processo que seu nome denota. No entanto, como visto em Baker (1995), corpus comparável é utilizado para análise de fenômenos no texto traduzido e é construído com base em originais em língua A e traduções para a língua A (mesma língua). O objetivo desta investigação recai também sobre a validação dos dados obtidos a partir da extração terminológica realizada no corpus paralelo, e não sobre busca por fenômenos existentes no texto traduzido.

2.4.10 Corpora Paralelos

Por paralelo entende-se um conjunto de itens (dados linguísticos) confrontados com outro conjunto de itens (dados

linguísticos relacionados). No caso de corpora paralelos, dispõe-se um item paralelo a outro. Tal justaposição pode se referir ao conjunto completo dos textos constituintes dos corpora ou a parcelas. Pode se referir, pois, a resultados de escalonamentos que abrangeriam, como posto, ou o conjunto dos textos, seguindo o paralelismo contido na segmentação de cada frase ou sentença, até o estabelecimento de parcelas unitárias, na relação do tipo: *entre palavras*.

Para Baker (Ibid.), um corpus/corpora paralelo é constituído de um original, que é considerado como texto fonte em língua A, acompanhadas de suas respectivas versões traduzidas, consideradas como texto(s) alvo(s) em língua B (outra língua). Baker (Ibid.) destaca a importância do papel dos corpora paralelos e sua aplicação no campo disciplinar da tradução, integrado aos programas de auxílio tradutológico. Nota-se também que corpora paralelos constituem fontes para a geração de recursos linguísticos para pesquisas de diversos fenômenos relacionados à linguagem, como os utilizados nesta pesquisa em que os dados gerados têm por propósito sua utilização em sistema de tradução automática e em campos mais amplos como o da inteligência artificial.

Por corpus paralelo entende-se aqui um conjunto de textos alinhados de forma que combinem, de forma biunívoca, isto é, mutua, com suas respectivas traduções. Esses textos em combinação bilateral também são conhecidos como textos paralelos. O conceito de *bitextos* não será aplicado nesse contexto devido ao fato de que os documentos e suas traduções não serão mesclados durante o processo de alinhamento. Para a extração terminológica o uso de corpora paralelos propicia a geração de recursos linguísticos que contêm informações relevantes para a compilação de glossários e bancos terminológicos através de cálculos realizados por meio do alinhamento.

Corpora paralelos são fontes valiosas em várias aplicações, como por exemplo: o estudo de traduções desenvolvidas por máquinas; rodagem de traduções automáticas; criação de léxicos e terminologias bilíngues, entre outras atividades.

2.4.11 Corpora Comparáveis

Os corpora comparáveis se sobrepõem às limitações dos corpora paralelos, uma vez que as fontes originais, ou seja, textos monolíngues, são muito mais abundantes que os textos traduzidos, sem mencionar o fato de que, atualmente, se encontram fartamente disponíveis na Internet.

No entanto, por natureza, as buscas de traduções em textos comparáveis são muito mais desafiadoras do que nos textos paralelos. A construção de um corpus comparável requer atenção, enquanto a definição de corpus paralelo é bastante simples (texto original e tradução). A construção de um corpus comparável exige controle sobre a seleção dos textos fontes em ambas as línguas.

Definimos, no escopo desta investigação e segundo critérios estabelecidos por Tagnin (2005), corpora comparáveis como referentes a textos originais, similares em ambas às línguas, relativamente ao par tratado pelo tradutor, mesmo gênero, tipologia similar, extensão e período de publicação próximos.

Podemos citar como exemplos de corpora: O COMET – um Corpus Multilíngue para Ensino e Tradução. Trata-se de um corpus eletrônico que tem como objetivo servir de suporte a pesquisas linguísticas, principalmente nas áreas de tradução, terminologia e ensino de línguas. O COMET[12] é composto por três subcorpora: i) CORTEC (Corpus Técnico), corpora comparável que faz parte do projeto; ii) COMAPREND é um corpus multilíngue de aprendizes, e o iii) CORTRAD, um corpus paralelo no par linguístico inglês – português.

2.4.12 Corpus Dinâmico, Corpus Estático, Corpus de Amostragem e Corpus de Monitoramento

Corpus dinâmico, baseados nos argumentos de Berber Sardinha (2004), permite que textos sejam incluídos ou suprimidos do corpus, de acordo com as necessidades e os objetivos do pesquisador, justamente o que caracteriza o corpus monitor. O corpus monitor é utilizado para acompanhar mudanças e variações ocorridas na língua ao longo do tempo. Por isso permite que, diacronicamente, sejam acrescentadas novas informações linguísticas ao material já estocado.

Contrariamente ao corpus estático, cujo formato não aceita acréscimos ou remoções de textos, o dinâmico se caracteriza como corpus de amostragem. Para Berber Sardinha (2004), um corpus de amostragem (*sample corpus*) é composto por porções de textos ou variedades textuais, planejadas para constituir uma amostra finita da linguagem como um todo.

[12] http://www.fflch.usp.br/dlm/comet/consulta_cortec.html

2.4.13 Corpora de Estudo e de Referência

De acordo com Berber Sardinha (2004) um corpus de estudo é aquele que pretende desenvolver determinada pesquisa, enquanto um corpus de referência, também conhecido como 'corpus de controle', funciona como termo de comparação para a análise. O segundo é normalmente muito mais extenso do que o corpus de estudo. O objetivo é comparar o corpus de referência ao corpus de estudo e, por meio de cálculos estatísticos (qui-quadrado ou razão de verossimilhança), determinar as palavras cujas frequências no corpus de estudo forem significativamente maiores sendo comparadas ao corpus de referência que, a partir dos resultados obtidos por meio do cálculo estatístico, serão consideradas como palavras-chave.

Berber Sardinha (2004) aborda ainda outros aspectos do corpus, como por exemplo: a *direcionalidade* da tradução. Segundo sua taxonomia, ele pode ser unidirecional (vetor em um sentido), por exemplo: do inglês para o português, bidirecional (nas duas direções mutuamente), do inglês para o português e vice versa, ou multidirecional (várias direções), por exemplo: do inglês para português e para o francês. Importante sublinhar ainda a *temporalidade*: distinguindo entre sincrônico e diacrônico. O corpus sincrônico envolve períodos de tempo bem delimitados: 1999, 2000, 2001, 2006, por exemplo. Diferentemente, no corpus diacrônico verificam-se lapsos de tempo bem mais amplos: 1765, 1897, 2009, 2013, por exemplo. Finalmente, o 'modo, fenômeno que Berber Sardinha (Ibid.) classifica como: modo falado ou escrito. O corpus falado corresponde a transcrições da fala, enquanto o escrito diz respeito a textos redigidos.

2.5 CONCEITUALIZAÇÃO DE TERMINOLOGIA (TE)

A necessidade de manuseio de grandes quantidades de palavras em diversos idiomas, com vistas a atingir melhor entendimento de sua distribuição e organização, tem conduzido os pesquisadores a atuarem de forma multidisciplinar. No aprendizado de um novo idioma, por exemplo, o indivíduo pode se deparar com situações em que uma dada palavra não denota exatamente aquilo que havia sido inicialmente captado (aprendido). Tal fenômeno pode, muitas vezes, ser gerado pelo contexto em que tal unidade foi empregada. Isso se torna mais evidente na tradução de textos especializados. Em virtude disto, o estudo da língua especializada tem como função a redução das implicações que as

palavras polissêmicas podem provocar em um texto científico. Busca assim afastar eventuais ambiguidades que geralmente possam emergir quando se trata de considerar a língua de base.

Uma das áreas de estudo que procurar lidar com esse fenômeno é justamente a Terminologia. No entanto, há longos debates em relação à considerar se a Terminologia é uma disciplina ou uma prática linguística, ou ainda um instrumento de auxilio à tradução. Um dos grandes estudiosos da Terminologia, Sager (1990), destaca que a Terminologia não pode ser considerada com uma disciplina autônoma, isso resultaria do fato de ainda se tratar um campo relativamente novo, parcialmente desprovido de arcabouços teóricos suficiente para a "proclamação" da mesma como uma disciplina autônoma..

Dado a afirmação de Sager (Ibid.), levanta-se a dúvida interna sobre o que significaria o termo "disciplina autônoma", uma vez que o próprio campo dos Estudos da Tradução nasce quase que concomitantemente ao surgimento da Terminologia. Ademais, parece se tratar de um campo essencialmente inter e multidisciplinar, na mesma medida que a interpretação ou a tradução.

Além das proximidades relativas ao período de surgimento de ambas as áreas, estudos em terminologia e tradução têm fortemente crescido após os anos 90 e caminham para o reconhecimento como disciplina passíveis de serem trabalhadas "fora" da Linguística. Neste trabalho a terminologia é conceitualizada por Cabré e Sager (1998/1999) como sendo uma "pratica bem desenvolvida" e principalmente em consonância com os objetivos desta pesquisa, que é justamente ao de servir como um instrumento de auxílio a intérpretes e tradutores.

Ao lançar sua definição de terminologia, Sager (1990, p.3) retrata o termo partindo de suas raízes etimológicas, como sendo um "termo polissêmico impróprio" e cita como exemplo uma palavra que possui diversos sentidos. Ainda ao considerar fatores etimológicos, o autor destaca que o próprio termo "terminologia" seria definido como uma "ciência/estudo/conhecimento dos termos" que faz um paralelo com a lexicologia, que tem por objetivo o estudo do léxico. Em um contexto histórico, e de modo breve, a terminologia se refere a um vocabulário técnico.

Dias (2000), em suas argumentações, traça uma discussão sobre a terminologia, sendo encarada como objeto e também como disciplina. Ao se basear em Cabré, Dias (Ibid.) ressaltam-se as três diferenças nas concepções sobre a ciência que trata dos termos. Segundo a autora, as três definições giram em torno de grandes áreas, a saber, a linguística, a filosofia e, por fim, as outras disciplinas técnico-científicas:

> Para a linguística, os termos são o conjunto de signos linguísticos que constituem um subconjunto dentro do componente léxico da gramática de determinada pessoa. Os termos, para a linguística, são uma forma de saber. Para a filosofia, a terminologia é um conjunto de unidades cognitivas que representam o conhecimento especializado. É, portanto, uma forma de conhecer. E, por fim, para as diferentes disciplinas técnico-científicas, a terminologia é o conjunto das unidades de expressão e comunicação que permitem transferir o pensamento especializado. Portanto, é uma forma de transferir, de comunicar. (DIAS, 2000, p. 90)

Dias (Ibid.) também realça o fato de a terminologia ser tratada como disciplina autônoma por alguns estudiosos e, por outros, ser defendida como ciência "autônoma e autossuficiente", principalmente por possuir fundamentações próprias mesmo não se desligando das bases de outras disciplinas afins, como a Linguística.

Dias (ibib.), ainda argumenta que dentre os estudiosos dessa ciência, os mais destacados são Sonneveld e Cabré, e que, com base no primeiro, reproduz que a "disciplina terminologia congrega conhecimentos oriundos de diferentes campos do conhecimento, tal como a informática (engenharia do conhecimento e inteligência artificial), a linguística (semântica, lexicologia e tradução), as ciências da documentação e classificação, a conceptologia e a nomenclatura", diálogos e integrações suficientes para dar origem ao nascimento de um novo campo de estudos "multidisciplinar com métodos e princípios próprios". Acrescente-se: domínio cujos resultados gerados vêm sendo disputado por pesquisadores que procuram respaldos científicos para suas afirmações. Aliás, uma das premissas básicas de toda e qualquer pesquisa científica moderna é afastar certezas e conclusões axiomáticas. Por outro lado, a orientação é buscar o máximo de pontualidade e precisão para basear quaisquer considerações. Talvez esta seja uma das características mais atraentes da Linguística de Corpus: a localização científica dos fatos da língua permeada por cálculos matemáticos.

Levando-se em consideração a leitura de Dias (Ibid.) a respeito dos argumentos de Sonneveld para o tratamento da terminologia

enquanto disciplina, pode-se remeter o leitor às base da criação do campo disciplinar dos Estudos da Tradução, já trazida à luz nas palavras de Holmes, e evocadas nas seções anteriores, através das quais se percebe que a tradução, para Sonneveld, se enquadra nas circunscrições do campo disciplinar da terminologia. O autor, citado aqui em *apud*, ainda pauta os Estudos da Tradução como disciplina que busca autonomia método-epistemológica. Mesmo que possua caráter interdisciplinar e multidisciplinar, ainda enfrenta dificuldades para definir acordos teóricos, diferentes da Terminologia. Esta última, talvez por suas tradições mais longas, define melhor seus raios de ação. Por tais razões, embora se possa contar com definições plurais, não são poucos os autores e organismos que se lançam à definição da referida ciência. A *International Association of Terminology*, por exemplo, oferece a seguinte descrição:

> Terminologia trata do estudo e utilização dos sistemas de símbolos e signos linguísticos utilizados para a comunicação humana em áreas especializadas de conhecimento e atividades. É principalmente uma disciplina linguística - a linguística sendo interpretada aqui no seu sentido mais amplo possível - com ênfase na semântica (sistemas de significados e conceitos) e pragmática. É interdisciplinar, no sentido de que ela também toma emprestado conceitos e métodos de semiótica, epistemologia, classificação, etc e está intimamente ligada as áreas cujos léxicos se descrevem e para o qual se destina a prestar assistência no ordenamento e uso de designações. Embora a terminologia tenha sido no passado mais ligadas aos aspectos lexicais das línguas de especialidade, o seu escopo se estende a sintaxe e fonologia. No seu aspecto aplicado, a terminologia está relacionada com a lexicografia e utiliza técnicas da ciência da informação e da tecnologia. International Association of Terminology (Sager, 1990, p.4, tradução nossa)[13]

[13] Terminology is concerned with the study and use of the systems of symbols and linguistic signs employed for human communication in specialised areas of knowledge and activities. It is primarily a linguistic discipline - linguistics being interpreted here in its widest possible

Como se constata, na definição acima apresentada, a terminologia é definida como campo interdisciplinar, mas não é necessariamente reconhecida como disciplina autônoma, sendo tomada como subárea de estudo no escopo mais geral da Linguística. Não se tem a intenção de, nesse breve trabalho, colocar em discussão o *status* da terminologia como disciplina autônoma ou não, mas tão somente definir seus principais conceitos e aplicações para a formação de uma proposta de extração terminológica fundamentada cientificamente.

Com base em Cabré (2003), ainda que a terminologia venha sendo estudada há muito tempo, a preocupação com o fenômeno da ambiguidade ganhou espaço apenas na primeira metade do século XX, sobretudo a partir dos estudos realizados pelo engenheiro austríaco Eugene Wüster (1898-1977), que postulou o objetivo principal da terminologia; qual seja: evitar ou contornar os fenômenos de ambiguidade na comunicação interlingual entre profissionais e especialistas de campos precisos. Ora, sabe-se que no escopo de um campo semântico ou lexical os fenômenos de polissemia podem se reduzir bruscamente. As pressuposições, que geralmente definem a elaboração de significações (locais, das palavras) e sentidos (geral, do texto) se orientam em razão das "esperas". Em outros termos, os papeis semânticos dos componentes regentes ou regidos assumirão os traços e componentes indicados por meio das orientações circunscricionais. Eis um exemplo: o conceito *paciente*, regido pelo verbo *operar* flexionado no particípio (e. g. O *paciente operado pelo médico*) dificilmente assumirá os traços de um <*agente caracterizado pela tranquilidade, que sabe aguardar sem alterar seus ânimos*>.

Como não poderia deixar de ser, Cabré (1999) também define também oferece seu conceito de terminologia, a ser aqui considerado de forma importante:

---

sense - with emphasis on semantics (systems of meanings and concepts) and pragmatics. It is inter-disciplinary in the sense that it also borrows concepts and methods from semiotics, epistemology, classification, etc. It is closely linked to the subject fields whose léxica it describes and for which it seeks to provide assistance in the ordering and use of designations. Although terminology has been in the past mostly concerned with the lexical aspects of specialised languages, its scope extends to syntax and phonology. In its applied aspect terminology is related to lexicography and uses techniques of information science and technology.

> Terminologia foi definida como: o processo de elaborar, descrever, processar e apresentar os termos de áreas de especialidade em uma ou mais línguas; a terminologia não é um fim em si, mas atende às necessidades sociais e tenta otimizar a comunicação entre especialistas e profissionais, fornecendo assistência direta a tradutores ou para comissões envolvidas com a padronização de uma língua. (Cabré, 1999, p.10, tradução nossa)[14]

O campo de estudo da terminologia surge com base na Teoria Geral da Terminologia, proposta pelo engenheiro austríaco Eugene Wüster. Sua principal motivação consistia em promover o desenvolvimento de um campo de estudo que tivesse como finalidade abordar o debate referente ao uso de léxico especializado em contextos específicos. Dentre os princípios propostos e desenvolvidos por Wüster, destacam-se os estudos e a padronização dos termos em áreas especificas do conhecimento.

De Wüster ao início da segunda década do século XX, a necessidade de se pensar sobre terminologia só fez ampliar-se, sobretudo após o advento da globalização e dos desenvolvimentos científicos, diga-se bastante acelerados, que se apresentam na área da microinformática oferecida ao grande público, aumentando ainda mais a demanda por traduções técnicas em razão das especificidades dos produtos virtuais oferecidos: vitrines, vendas, buscas. Tal processo não poderia deixar de desembocar na busca por materiais de apoio à tradução de volumes de textos incalculáveis e, por conseguinte, de suportes adequados à prática dos tradutores, sejam eles profissionais ou *free lancer*, se não contornarmos as realidades. Neste sentido, torna-se pertinente lembrar a seguinte observação de Krieger e Finatto (2004):

> Neste contexto de alargamento de fronteiras e de grande ampliação de intercâmbios, as línguas passaram a entrar mais fortemente em contato,

[14] Terminology was 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.

> exigindo novas competências linguísticas, em que se inclui o domínio dos termos técnicos. Junto a essas novas necessidades encontra-se a crescente demanda pelas traduções técnicas, as quais necessitam transpor adequadamente as terminologias de uma língua para outra. (KRIEGER e FINATTO, 2004, p. 18).

O desenvolvimento da terminologia evoluiu proporcionalmente aos avanços e expansões da (micro)informática. Hoje não se imagina mais que os primeiros computadores chegaram a ocupar 10 metros quadrados, que os mais avançados nos anos 1960-70 ainda eram manipulados em DOS[15], tampouco que usavam cartelas perfuradas, em papel. Outrossim, eram restritos a certos usos e usuários, e geralmente instalados em empresas de grande porte. Seria inconcebível possuir uma máquina desse tipo em uma residência normal, bem como pouco provável que o consumidor médio possuísse recursos para adquirir tais equipamentos.

A partir dos anos 1980, quando ferramentas informatizadas começaram a se popularizar, os softwares passaram a permitir o processamento de dados em quantidades cada vez maiores, favorecendo a criação de bancos de dados locais e, consequentemente, a utilização dos corpora em ambientes domésticos e acadêmicos setorizados. A partir dos anos 1980 surgiram os primeiros dicionários digitais a venda no mercado. Inicialmente, em termos de conteúdos, se basearam primordialmente sobre os materiais já existentes nas versões em papel, mas com as vantagens das buscas automáticas, permitindo acessar as entradas de modo instantâneo, de verificar as informações referentes ao léxico de forma categorizada: etimologia, fonética, denotações, associações, exemplos, antônimos. Breve, não há como negar o imenso salto para o futuro que se operava naqueles anos.

Embora existam vários tipos de dicionários, ao se deparar com um texto de especialidade, o tradutor nem sempre encontra o suporte necessário para a realização de seu trabalho. Ainda são raros os bancos de dados capazes de responder a todos os domínios implicados na

[15] O DOS, sigla para Disk Operating System ou sistema operacional em disco1 é um acrónimo para vários sistemas operativos intimamente relacionados que dominaram o mercado para compatíveis IBM PC entre 1981 e 1995, ou até cerca de 2000 caso sejam incluídas as versões de Microsoft Windows parcialmente baseadas em DOS 95, 98 e Me. (http://pt.wikipedia.org/wiki/DOS).

tradução. Nesse sentido, as ferramentas computacionais têm auxiliado na composição desses materiais de consulta que, progressivamente, geram novas perspectivas nas atividades de busca por correspondentes tradutórios.

A maioria dos estudiosos da Linguística de Corpus, citados no escopo dessa tese, unanimemente, destacam o momento histórico no desenvolvimento de estudos relacionados à área da terminologia apontando os princípios desenvolvidos por Wüster como sendo a: primeira teoria de Terminologia. Trata-se da Teoria Geral da Terminologia (TGT), baseada nos estudos de Wüster que concebe o termo como um rótulo designativo de uma unidade de conhecimento, desconsiderando sua dimensão linguística. Tal posicionamento está relacionado ao fato de que, historicamente, as terminologias correspondiam às nomenclaturas científicas, utilizadas largamente pelas ciências taxonômicas, tal como a química e a botânica (cf. Krieger, 2008, p. 5).

Com o desenvolvimento de estudos baseados naqueles propósitos outrora apresentados por Wüster, naturalmente lhe foram dirigidas muitas críticas. Esse processo deu início a um novo campo de estudo referente à terminologia, chamado de *Teoria Comunicativa da Terminologia*. Proposta por Cabré em 1993, a referida teoria tinha por objetivo estabelecer um caráter de maior flexibilidade dos processos comunicativos em contextos específicos. Com o argumento de que Wüster não fornecia aparato teórico para a descrição do léxico específico, em relação a esse novo campo de estudo, encontra-se a ponderação de Almeida, que nos parecer ser aqui pertinente:

> [...] a TGT começa a dar lugar a uma teoria mais ampla e flexível, denominada Teoria Comunicativa da Terminologia (TCT), cujo instrumento teórico-metodológico pode explicar melhor os fenômenos que envolvem a comunicação especializada e melhor descrever suas unidades mais representativas, os termos, de forma a abranger toda sua complexidade. (ALMEIDA, 2003, p. 22).

A área da terminologia constitui um campo de estudo passível d ser divido em três subáreas, a saber: a terminologia, a lexicografia e a

lexicologia. Outro elemento relacionado ao campo de estudos da Terminologia que parece de extrema importância no presente escopo são os subcampos relacionados, formando a grande área dos estudos lexicais, que podem ser mais bem definidos:

> Terminologia já escreveu um percurso histórico que tem impulsionado sua investigação de forma a alinhá-la à Lexicologia e à Lexicografia, formando o trio de disciplinas que configuram as denominadas Ciências do Léxico. (KRIEGER, 2011, p. 445)

Baseados no argumento de Krieger e Finatto (2004), o escopo geral da terminologia teria por objetivo estudar o tratamento concedido aos termos técnicos e/ou científicos. A terminologia focaria então o "conjunto de termos técnico-científicos, representando o conjunto das unidades lexicais típicas de uma área científica, técnica ou tecnológica". Como já salientado, o foco desta investigação centra-se prioritariamente em material textual de um campo científico específico: as ciências médicas.

Fato de grande importância a se destacar ao se discutir sobre o status da Terminologia enquanto ciência concerne à argumentação de Krieger (2011). Para a pesquisadora, especialista em estudos terminológicos e também em tradução, o termo pode ter significados diferentes. Segundo ela, escrito com a inicial da palavra em caixa baixa (minúscula) a palavra poderá significar:

> [...] conjunto de termos de áreas científicas, técnicas, tecnológicas, a exemplo da terminologia médica, da economia, do direito, da linguística etc. Esta é uma forma também de dizer que não há conhecimento especializado sem seus termos próprios, o que remete à relação direta com os conceitos de cada campo de saber especializado. KRIEGER, 2011, p. 445)

Ainda se pode destacar, em conformidade com as considerações da referida autora que, há muito, existem termos próprios ligados a áreas

específicas. Em outras palavras: léxicos especializados. Corroborando com as observações de Krieger, como já observado, a necessidade de fazer, cada vez mais, circular a informação, a urgência e premência das demandas dos mercados se desenvolvendo concomitantemente à geração de novos recursos de naturezas plurais, desemboca na necessidade de processamento de dados linguísticos em tempo hábil. Instalam-se efeitos cascata em progresso, ou seja, evolução terminológica ligada ao aprimoramento das áreas do saber, evolução dos meios para o processamento das línguas. A explosão das bases terminológicas não se desenvolve *intramuros* acadêmicos. Trata-se, já há algumas décadas, de uma questão política e de mercados (leia-se: financeira).

Quanto ao outro tipo de registro (grafia) da palavra *Terminologia*, com sua inicial em caixa alta (maiúscula),Krieger observa que essa fórmula se refere ao campo disciplinar da própria Terminologia. Para a autora, as duas formas, caracterizadas pela diferenciação da notação entre as duas palavras é bastante recente e passou a ser utilizada em meados do século XX, demonstrando também o quão tenro ainda é o campo disciplinar da Terminologia no meio acadêmico, apesar de sua longa tradição ao derivar para os estudos etimológicos. Como explicitado aqui, nas linhas desse trabalho, utilizaremos *Terminologia* (com inicial em maiúscula) para definir o campo teórico, enquanto *terminologia* (com inicial em minúscula) quando tratarmos de *terminologia* específica referente a áreas com aquela investigada nessa pesquisa, ou seja, a *terminologia* das ciências médicas, ou de forma extensiva, como *terminologias* de quaisquer outras áreas: jurídica, agrícola, astronômica, etc.

Ainda tratando a terminologia, Krieger (2011) destaca um fenômeno que ocorre em certas áreas do conhecimento, sendo pertinente em relação o foco desse trabalho. De modo indireto, em consonância com o objeto de pesquisa aqui tratado, a autora evidencia a grande precisão passível de ser alcançada em relação aos resultados obtidos por meio do processamento automático que, em muitos sentidos, ultrapassa a esfera dos cálculos estatísticos:

> Diferentemente, acontece com os termos de algumas áreas científicas que, em seus processos denominativos, inspiraram-se na tradição das nomenclaturas técnico-científicas, como as da Biologia e da Zoologia, que foram cunhadas em

> latim e em grego com a finalidade de fugir das ambiguidades a que o léxico comum está sujeito. O princípio constitutivo das nomenclaturas, caracterizando-se como uma espécie de língua universal das ciências e das técnicas, influenciou largamente o processo denominativo de muitas áreas científicas. (KRIEGER, 2011, p. 444)

Como a autora clarifica em seu discurso, há um princípio constitutivo aplicável às nomenclaturas que, em muitos casos, se tornam procedimentos universais. Tais ações se justificam no sentido, por exemplo, de evitar ambiguidades em campos disciplinares tão específicos. Assim como este trabalho tem sua base de desenvolvimento situada na extração de itens lexicais e definição de correspondentes tradutório em um campo que interfere nos processos de constituição e de mudança nas línguas modernas, sua aplicação em campos como o das ciências médicas poderá, eventualmente, ser pertinente, sobretudo em se considerando que o material investigado no âmbito desta pesquisa não se restringe apenas a uma área em específica, se estendendo aos diversos campos do saber anexos e, por vezes, dependente: tal como o Direito Médico, a Medicina Legal, integrando, por exemplo: Medicina + Jurídico.

Com relação à tipologia textual utilizada como insumo para as análises propostas nas páginas que seguem, e ainda relativamente às ciências médicas, Krieger destaca os formantes da terminologia empregada nessa área. Ponto importante de se realçar é a proximidade de algumas línguas de base, por exemplo, como o latim e com o grego. Poder-se-ia ainda mencionar o hebraico. Sobre a terminologia empregada nas ciências médicas, faz-se importante remeter o leitor ao excerto abaixo:

> É assim que termos da Medicina, por exemplo, utilizam formantes gregos e latinos em seus repertórios terminológicos, como atestam litíase renal e cardiopatia. Tal formato terminológico, mesmo obedecendo aos padrões morfossintáticos do português, tende a uma exclusividade designativa, restringe-se ao estatuto de termo, já que não passou a ser usada em conversas informais. (KRIEGER, 2011, p. 444)

Naturalmente, as especificidades da terminologia se estendem muito além de definições de itens lexicais ligados a determinadas áreas do saber. Não é raro que a terminologia seja trabalhada para representar, e até mesmo para padronizar recursos das línguas em certos ambientes nos quais ainda persistem indefinições ou pluralidades referenciais. Sua aplicação é de extrema relevância uma vez que a apresentação, manutenção e desenvolvimentos de diversos aplicativos e produtos estão atrelados a manifestação linguística, gerando implicações comerciais e científicas, tal como já aventado acima. Com base na Figura 4, apresentada pelo *Pointer Final Report*[16], pode se ter uma noção breve de algumas das áreas ligadas à terminologia.

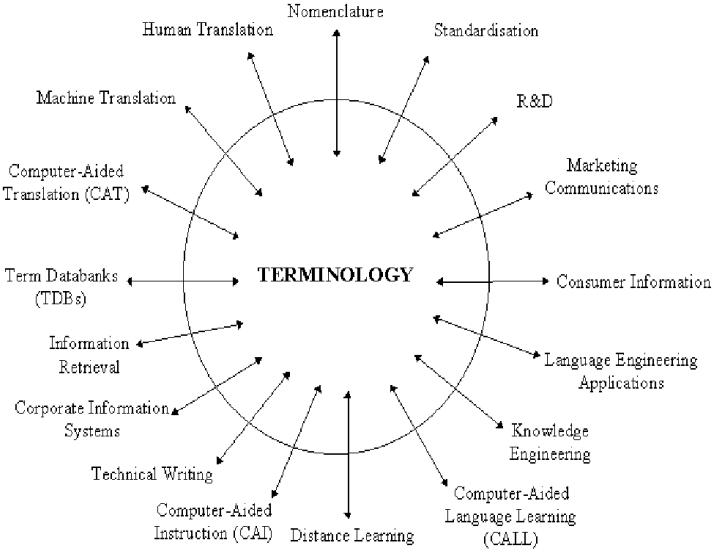

Figura 4: Produtos relacionados à terminologia (DIT, 1998)

[16] The POINTER project consortium comprises a representative cross-section of organisations involved in terminology use and in the development and dissemination of terminology resources. The project partners include public and commercial organisations from Austria, Belgium, France, Germany, Greece, Italy, Scandinavia, Spain, Switzerland, The Netherlands and the United Kingdom. http://www.computing.surrey.ac.uk/ai/pointer/

Em Silva e Luz (2004), a terminologia enquanto ferramenta tem servido como fonte de subsídios em diversas áreas do conhecimento. Entre as áreas citadas pelos autores, destacamos aqui:

> [...] formulação e sistematização do conhecimento (classificação conceitual para cada disciplina científica); transferência de conhecimentos, habilidades, experiências e tecnologia; tradução de termos científicos; elaboração de resumos de informações. (SILVA e LUZ, 2004, p. 2)

Centramos o foco particularmente sobre o que Silva e Luz (Ibid.) pautam como "tradução de termos científicos". Alvo desta investigação, a terminologia é considerada aqui como ferramenta para a resolução parcial de questões que eventualmente se apresentam durante a atividade tradutória, uma vez que os tradutores, de forma geral, parece que se colocam, cada vez mais, face às diversas áreas do conhecimento, uma vez que, nas últimas décadas, dão a impressão de se desdobrar em proporções geométricas. Sendo assim, torna-se praticamente impossível acreditar que um tradutor juramentado, por exemplo, esteja apto a traduzir quaisquer tipos de textos, expressos em quaisquer áreas do conhecimento.

Acreditando que os estudos terminológicos possam exercer o papel de agentes de instrução para o tradutor, pode-se conceber, assim como já observado por Dias (Ibid.) que a terminologia se apresenta como um recurso "técnico-científico" quando a tratada como instrumento de suporte e auxilio ao exercício tradutológico. Dias (Ibid.) ainda afirma que sem a utilização dessa ferramenta, isto é: a terminologia, os especialistas poderiam não conseguir realizar seus trabalhos com êxito, uma vez que a comunicação poderia ficar limitada, dificultando a tramitação nas trocas e exposição do conhecimento referente a determinada área do saber. A apresentação organizada, normatizada, compartilhada, sistematizada, está implícita às ciências.

Silva e Luz (2004) corroboram com a colocação de Dias (Ibid.) destacando a importância da terminologia técnica, principalmente quando relacionada à tradução. Para eles:

> Diversos são os casos de ausência de comunicação pela não observância da

> terminologia técnica, quando o conceito não está claro ou quando a tradução não é devidamente cuidadosa. (SILVA e LUZ, 2004, p. 4)

Ao se tratar de terminologia como definição de conceitos ou como ferramenta para catalogação de correspondentes tradutórios de um determinado texto de especialidade, percebeu-se a necessidade de ferramentas com alto nível de desempenho e precisão para extração, recuperação e catalogação, como também a clara visão da importância que a terminologia exerce no escopo da comunicação circunscrita e pontual.

2.6 TERMINOLOGIA AD HOC

Na pratica tradutória efetiva, em geral o tradutor necessita, na maioria dos casos em que recorre aos glossários e aos bancos terminológicos, é encontrar o considerado equivalente (equivalente usado nesse trabalho se refere ao correspondente tradutório adequado) capaz de estabelecer elos pertinentes entre as línguas confrontadas. Essa busca pelo correspondente do termo em ambas as línguas é conhecida por *ad hoc*[17].

Para a realização dessa busca, os tradutores geralmente recorrem a bancos de dados linguísticos, glossários. Por vezes, também realizam buscas até mesmo em bancos de textos que já traduzidos por eles. Em alguns casos, ao não encontrar o correspondente adequado, trata-se de recorrer a textos similares àqueles que estejam traduzindo. Um dos exemplos concerne aos certificados, contratos, acordos, pareceres, sentenças, etc. Em geral, tais confirmações, baseadas em dados empíricos podem gerar excelentes resultados.

2.7 EXTRAÇÃO TERMINOLÓGICA

As necessidades que conduzem, como resposta, à construção de glossários e bancos terminológicos têm se intensificado, sobretudo em razão das pesquisas *on-line*. Para o desenvolvimento da tarefa, as investigações derivam para o desenvolvimento de ferramentas que

[17] Ad hoc é uma expressão latina cuja tradução literal é "para isto" ou "para esta finalidade". http://pt.wikipedia.org/wiki/Ad_hoc

possibilitem a identificação automática de equivalentes tradutórios. Esta perspectiva faz parte de uma nova geração de produtos que, por extensão, geram novas abordagens para a extração terminológica.

Neste início do século XXI, Jacquemin (2001) se dedicou à proposta de identificação automática para candidatos a termos em um dado conjunto de material textual. Porém, deve-se levar em conta que após a apresentação de seus trabalhos, já decorreram 12 anos, ou seja, tempo suficiente para que os recursos informáticos conduzissem a novos horizontes. No campo da informática, assim como em campos científicos em evolução constante, lapsos superiores a 10 anos deverão ser observados com bastante atenção. Se uma década é capaz de provocar variações e mudanças na língua geral, pode-se, por extensão, deduzir que as metamorfoses que incidem sobre técnicas e produtos induzirão à novas fórmulas de expressão. Fenômenos desta ordem não acontecem somente no âmbito de um sistema linguístico, mas opera em consonância com as alterações que ocorrem nos idiomas em que determinados campos mais se desenvolvem. No caso do português brasileiro, algumas áreas sofrem influências do inglês (cf. Informática, outras do francês (Materiais), outras do alemão (Filosofia). Há até mesmo áreas, como a Biologia, a Astronomia, que continuam gerando termos com base no latim e no grego. Por isso, a consideração de outros idiomas se mantém como fator de referência.

Para Kageura et al. (2004), as ferramentas desenvolvidas para tratar da extração terminológica deveriam abordar diversas questões linguísticas, baseadas em estatística e considerando regras e métricas em conjunto para que fosse possível extrair, a partir dela, listas de candidatos a termos, justamente daqueles mais salientes à construção de glossários e bancos terminológicos. Para o processamento do conjunto de materiais textuais, algumas regras e métricas de cunho linguístico têm sido incorporadas aos sistemas. Dentre elas, citamos a anotação de Classes de Palavras (*POS – Part-of-Speech*) e a Lematização. Essa nova abordagem, em conjunto com as ferramentas fornecidas pela Linguística de Corpus, tornam o nível de precisão, durante o processamento e identificação dos candidatos a termos, bastante apurado.

Como em todo processo automático de anotação de corpora, questões relacionadas aos eventuais *erros* (leia-se *inadequações*) que emergem do processo são supervisionados, tal como a anotação e a geração de candidatos a termos. Para este trabalho, optou-se por usar modelos da língua para anotação de classes de palavras e lemas de maneira não supervisionada uma vez que o conteúdo dos corpora compreendem aproximadamente 30 milhões de palavras, tornando a

tarefa inviável. Já com relação ao alinhamento, para garantir maior nível de acerto durante o processo, ao nível de sentença e da palavra, o corpus foi revisado, e assim fornecendo o alinhamento adequado a cada segmento.

Tendo em vista que a atividade de extração terminológica se opera por meio de corpus paralelo, os fatores considerados para a extração terminológica serão os seguintes:

(a) geração de palavras-chave para os textos em língua inglesa;

(b) alinhamento estatístico nível de palavras entre os textos em língua inglesa e língua portuguesa para geração de unidades lexicais probabilísticas para tradução;

(c) recuperação das informações entre as listas de palavras-chave e a relação de tradução probabilística extraída do alinhamento;

(d) geração da lista final com os candidatos a termos e suas possíveis traduções, baseadas no processo estatístico de alinhamento.

As discussões contemplam a extração terminológica de forma semiautomática, uma vez que conta com a interferência do pesquisador na tarefa de preparação do conjunto de materiais textuais e manipulação das etapas de aplicação durante o processamento. Tal procedimento responde à necessidade de testar a metodologia escolhida para o desenvolvimento futuro de uma ferramenta automatizada com vistas à automatização total das etapas semiautomáticas.

Com relação à extração terminológica, podemos entender na afirmação de Almeida (2007) que:

> A extração de termos diz respeito à obtenção do conjunto terminológico que comporá a nomenclatura 2 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 pode ser feita de forma manual ou automática, entretanto, quando se utiliza extração automática, é necessária a elaboração de córpus

> em formato digital, evidentemente. (ALMEIDA et al., 2007, p. 410)

Teline (2004), ao tratar da literatura da abordagem estatística, nos mostra o fato que há uma escassez de trabalhos recentes envolvendo sistemas ou mesmo algoritmos estatísticos relacionados à extração terminológica. A pesquisadora ainda chama atenção para a existência, dentre poucos, do trabalho aqui já discutido de Dias et. al. (2000) sobre a utilização de n-grama denominado Esperança Mútua e a proposta de *LocalMaxs* para a extração terminológica. No entanto, esses trabalhos não nos remetem à extração terminológica bilíngue por meio de máquina, fato que está sendo justamente abordado na presente pesquisa.

A pesquisadora trata dos vários métodos de extração automática de terminologia relativamente a textos em português. Apesar de seu foco consistir de conjuntos de textos monolíngues, Teline (2004) realiza a análise de um dos cálculos estatísticos utilizado em nossas análises, a saber: a razão de verossimilhança. O cálculo para a análise da razão de verossimilhança será apresentado, na sequência deste trabalho, na seção de Fundamentação Metodológica. Outrossim, explanaremos o processo de forma detalhada.

3 FUNDAMENTAÇÃO METODOLÓGICA

3.1 APERFEIÇOAMENTOS AOS SUPORTES INFORMÁTICOS

Face aos interesses em se procurar alcançar melhores resultados na recuperação dos candidatos a termos, propõe-se a utilização de três técnicas diferentes baseadas em cálculos para a obtenção das palavras-chave. Após o levantamento dos dados, se optará por aquela técnica que gerar os resultados mais refinados. Dentre os procedimentos de cálculo empregados atualmente, recorreu-se ao qui-quadrado de Pearson (Pearson, 1900), que tem por finalidade investigar as possibilidades de uma palavra de se destacar em determinados conjuntos de textos. A referida medida é realizada sobre os valores esperados e sobre aqueles observados, com o objetivo de verificar se o fenômeno pode ou não ter ocorrido acidentalmente, ou se determinado elemento revela-se como candidato a termo naqueles conjuntos de textos. O qui-quadrado de Pearson é desenvolvido através do seguinte cálculo, por nós examinado e reproduzido:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

A fórmula, de Pearson, aplica-se a valores universalmente absolutos, ou seja, como o conjunto de textos a ser analisado nesta investigação é representativo dos estudos realizados na área de medicina, publicados e traduzidos no âmbito da comunidade europeia, porém, e naturalmente, não em sua totalidade. O cálculo será realizado por meio da correção do qui-quadrado de Yates (Yates, 1934), que visa atingir a probabilidade entre uma tabela de contingência, considerando os valores como a proporção do que seria universal, como pode ser constatado na fórmula que segue abaixo:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Por fim, serão analisados os resultados obtidos com a aplicação de cálculo adicional, a saber: a *razão de verossimilhança* (cf. Wilks, 1938). O referido logaritmo visa à função de verossimilhança, calculada através dos valores atribuídos e estimados por meio dos coeficientes em questão e expressos pela seguinte fórmula:

$$-2\log\Lambda = 2\sum_{i,j} k_{ij}\log\frac{n_{ij}}{m_{ij}}.$$

Para que o cálculo fosse automatizado, foram desenvolvidos para esta pesquisa, com base na fórmula de *razão de verossimilhança* e também de *qui-quadrado* e a correção de *Yates* sobre o *qui-quadrado*, programas computacionais para utilização em corpora de amostragem em *JavaScript* e para execução dos corpora completos em *C++* para cada uma das fórmulas em que o processo de busca por candidatos a termos pudessem ser obtidos com rapidez e precisão, tendo em vista a proposta deste estudo, a ser novamente lembrada resumidamente, a saber: o desenvolvimento de percursos técnicos com vistas a viabilizar a criação de glossários bilíngues através de processos estatísticos. Os códigos compilados para a realização de cada um dos cálculos podem ser acompanhados abaixo de forma linear e, subjacentemente, paradigmático:

Qui-quadrado de Pearson:

```
function _x2(a,b,c,d){
        x2=((Math.pow(a-
(b*(a+c))/(b+d),2))/((b*(a+c))/(b+d)))+((Math.pow(c-
((d*(a+c))/(b+d)),2))/((d*(a+c))/(b+d)));
        return(x2);
}
```

Qui-quadrado de Yates:

```
function _yx2(a, b, c, d){
```

```
        yx2=((Math.pow((a-(b*(a+c))/(b+d)-
0.5),2))/((b*(a+c))/(b+d)))+((Math.pow((c-((d*(a+c))/(b+d))-
0.5),2))/((d*(a+c))/(b+d)));
        return(yx2);
}
```

Razão de Verossimilhança:

```
function _ll(a, b, c, d){
        ll=2*(a*Math.log(a)+b*Math.log(b)+c*Math.log(c)+d*Math.lo
g(d)-(a+b)*Math.log(a+b)-(a+c)*Math.log(a+c)-(b+d)*Math.log(b+d)-
(c+d)*Math.log(c+d)+(a+b+c+d)*Math.log(a+b+c+d));
        return(ll);
}
```

Para realizar a testagem das fórmulas traduzidas em linguagem computacional, apresentadas acima, com base na fórmula de Dunning (1993) referente ao cálculo de razão de verossimilhança, foram realizados cálculos e rodagens-pilotos[18] para a certificação de que os procedimentos de programação estariam corretos e os resultados obtidos em processamento em conformidade com a aplicação em questão.

A seguir, apresenta-se um cálculo demostrando os dados que seguem abaixo e para os quais deverão ser atribuídas as seguintes convenções:

a = número de ocorrências de determinada palavra, obtidas no corpus de estudo;

b = o total de palavras no corpus de estudo;

c = número de ocorrências da palavra obtida no corpus de referência e finalmente;

d = o total de palavras no corpus de referência.

De forma a exemplificar a equação desenvolvida para cada item lexical nos corpora estudados, o cálculo foi desenvolvido da seguinte

[18] Chamamos de teste piloto a etapa prévia de testagem dos produtos empregados na pesquisa, de modo a evitar falsas manipulações ou escolhas.

maneira nesta pesquisa, como parte de estudo piloto para que fosse possível se certificar da confiabilidade dos dados obtidos por meio dessa computação na fórmula convertida para lógica testada:

2(a*log(a) + b*log(b) + c*log(c) + d*log(d) | - (a+b)*log(a+b) - (a+c)*log(a+c) | - (b+d)*log(b+d) - (c+d)*log(c+d) | + (a+b+c+d)*log(a+b+c+d))*

Sendo:

a=411 | b=515 | c=242242 | d=1015483 | a+b=926 | c-a=241831 | d-b=1014968 | c+d-a-b=1256799 | E1 = c(a+b) / (c+d) | E1 = 242242 *(411+515) / (242242+1015483) | E1 = 242242 * 916 / 1257725 | E1 = 224316092 / 1257725 | E1 = 178,35066648114651454014192291638 | E2 = d*(a+b) / (c+d) | E2 = 1015483**(411+515) / (242242+1015483) | E2 = 1015483 * 926 / 1257725 | E2 = 940337258 / 1257725 | E2= 747,64933351885348545985807708362 | G2 = 2*((a*ln (a/E1)) + (b*ln (b/E2))) | G2 = 2*((411*ln (411/178,35066648114651454014192291638)) + (515*ln (515/747,64933351885348545985807708362))) | G2 = 2*((411*ln (2,30444891577372879695140195291 91)) + (515*ln (0,68882559899588706927530888072075))) | G2 = 2*(411 * 0,83484156577010491289459246573995) + (-0,16189072146138160191087292346138) | G2 = 2*(343,11988353151311919967750341912) + (-0,36190929195897938701292045403 62) | G2 = 686,2397670630262 + (-0,36190929195897938701292045403 62) | G2 = 685,8778577710672*

Os resultados foram comparados com outros programas de análise lexical, como por exemplo, o *WordSmith Tools* (Scott, 1996, 1997, 1999, 2004, 2008), com a finalidade de legitimar que a conversão das fórmulas em programas computacionais realizadas para este trabalho reproduzissem os resultados compatíveis com aqueles alcançados em programas de alta confiabilidade. Outrossim, que estes algoritmos de programação pudessem ser inclusos na proposta de ordem sequencial desta pesquisa, justamente com a finalidade de não depender de programas oferecidos por terceiros.

Outros estudos já foram realizados por meio destes processos estatísticos. Cita-se, como exemplo, o trabalho das frequências das palavras no par inglês americano/inglês britânico, realizado por Hofland e Johansson (1982), cujo foco centrou-se em observar as palavras mais típicas entre as variantes da língua inglesa ao usar o corpus de estudo *The Lancaster-Oslo/Bergen Corpus* (LOB) e o corpus de referência

Brown, baseando-se no teste de qui-quadrado. Rayson, Leech e Hodges (1997) empregou, igualmente, um subcorpus do BNC e o teste de qui-quadrado para suas análises. Ainda se pode citar Rayson e Garside (2000) que estudaram a comparação de corpora de controladores de voo em relação ao subcorpus falado do BNC, no qual aplicaram a estatística da razão de verossimilhança para o processamento dos corpora.

Com os candidatos a termos obtidos por meio dos cálculos acima apresentados, a última etapa desta proposta de processo remete justamente ao objeto visado nesta investigação. O trabalho tomará, com efeito, técnicas já desenvolvidas para situações em que os pesquisadores se vejam eventualmente limitados, ou cerceados pelo tempo em suas tarefas de recuperar correspondentes de forma manual. Neste sentido, a proposta buscará mecanismos que se pretendem *eficazes* para a busca por correspondentes tradutórios de forma automatizada. Precisamente, nesta etapa inicial, todo o procedimento metodológico empregado deverá ser revisto, no sentido de incitar adaptações e, por conseguinte, reconstrução de novos modelos com a capacidade de propiciar a recuperação de correspondentes tradutórios com nível de precisão aceitável, de forma a garantir que cada entrada do glossário esteja respaldada ao máximo, mesmo que esteja sendo gerada de forma automática, diferente daquela que se poderia fazer com o trabalho qualitativo pontual, decorrente de apreciações e juízos humanas.

Como se observou na literatura específica, os dados utilizados e recuperados por meio do uso de ferramentas de processamento de corpora possuem ligação estreita com os trabalhos desenvolvidos na área da lexicometria, da lexicografia e da lexicologia. Tarefas outrora realizadas manualmente passaram a ser realizadas com maior precisão e rapidez. Naturalmente, sabe-se que as abordagens de cunho quantitativo antecedem a disciplina em questão, mas não se pode negar, como observa Villas Boas (2009), abaixo citado, que as visões em relação à localização de dados textuais sofreram profundas transformações, sobretudo após o advento da microinformática doméstica (i.e. voltada ao grande público). Veja-se:

> Dados empíricos têm sido usados em lexicografia muito antes da disciplina de linguística de corpus ser inventada. Os corpora, no entanto, mudaram o modo pelo qual os linguistas podem examinar uma língua. (VILLAS BOAS, 2009, p. 9).

Essa função, creditada ao uso de corpora, cresceu de forma considerável nas duas últimas décadas, acompanhando a revolução paralelamente gerada nas áreas em que o processamento automático de línguas faz parte das ações de primeira ordem. A título de exemplificação, remete-se o leitor à intensificação do emprego de corpora no ramo da lexicografia. Neste sentido, pode-se observar que:

> [...] para os lexicógrafos, sobretudo, foi uma revolução. Máquinas mais potentes e programas mais eficientes criaram para o lexicógrafo essa prancheta inestimável para sua arte de construir dicionários. O corpus veio possibilitar um confronto entre a teoria e os dados empíricos da língua. De fato, o corpus pode mostrar como funciona uma língua natural em escala reduzida. (BARROS, 2004, p. 264).

Não caberia aqui nestas páginas, nem seria possível, listar exaustivamente os pesquisadores que têm abordado o uso de corpora nos estudos lexicográficos. Todavia, embora ultrapasse os objetivos desta investigação, faz-se importante ainda destacar Campion e Elley (1971), Praninskas (1972), Kilgarriff, (2000), cientistas cujos trabalhos se aproximam dos objetivos aqui visados, mesmo se com relação aos dois primeiros trate-se de investigações realizadas com parcos recursos, naturalmente se comparados às performances oferecidas pelos processadores o século XXI.

Com destaque no grande impulso tomado na área da lexicografia e da terminologia, com a integração da adoção de ferramentas informatizadas para o processamento de corpora para o desenvolvimento de pesquisas, acredita-se que, ao fazer uso da metodologia aqui apresentada, será possível vislumbrar alguns passos adiante nas aplicações desenhadas – pretendida e humildemente – como eventuais suportes às atividades envolvendo a tradução.

4 METODOLOGIA: PROJETO, CONSTRUÇÃO E PROCESSAMENTO DOS CORPORA

O desenvolvimento de trabalhos apoiados na metodologia de *corpora* para pesquisas linguísticas e tradutológicas desenvolvidas apresenta, geralmente, três etapas, *viz*:

(i) projeto;

(ii) construção; e

(iii) processamento.

A presente investigação, neste aspecto, não ousa trazer nenhuma modificação de porte considerável. Logo, segue-se o modelo comumente utilizado por outros pesquisadores, buscando, todavia, cada vez que possível, aperfeiçoar as ferramentas e adequar a proposta de ordem sequencial com a finalidade de extração terminológica bilíngue automatizada.

4.1 PROJETO DO CORPUS

Trata-se da etapa inicial para o desenvolvimento de uma metodologia baseada em *corpus*. Para a maioria dos autores consultados, nessa fase é possível delinear o planejamento do que se deseja e espera alcançar ao compilar um *corpus*. O projeto do *corpus* constitui o marco inicial para as decisões a serem tomadas relativamente ao objetivo de criação, conteúdo, representatividade, tipo e direitos autorais (cf. Baker, 1995).

Como pôde ser visto em outras propostas de metodologias baseadas em *corpus*, já referenciadas acima, o planejamento do *corpus* deve ser definido com base em diversos critérios para não haver frustrações no desenvolvimento do trabalho. Estes critérios, em princípio, devem proporcionar alinhamento adequado ao objetivo da criação do *corpus* e, sobretudo, aos objetivos gerais da pesquisa. Essa etapa é dividida em cinco seções, a saber:

(i) objetivo da criação do *corpus*;

(ii) tipo do *corpus*;

(iii) seleção dos textos;

(iv) representatividade; e

(v) direitos autorais.

4.1.1 Objetivo da Criação do Corpus

No processo de criação do *corpus*, o passo inicial a ser tomado consiste em estabelecer o objetivo que conduziram a sua construção. Para a definição desse objetivo, há de se considerar que ele deva, em princípio, estar primeiramente alinhado àquela proposta explicitada no objetivo da própria investigação. A obtenção dos dados almejados para as análises, assim com os resultados da pesquisa, se basearão fortemente nessa definição-chave.

Alinhado ao objetivo da pesquisa, o objetivo de criação do *corpus* se resume em realizar a extração terminológica e investigar elementos metodológicos para a criação do banco terminológico bilíngue na área almejada, ou seja, das ciências médicas.

Finalmente, preconiza-se o recurso às ferramentas computacionais para a realização da extração terminológica, geração de lista de palavras, listas de palavras-chave e construção de dicionários probabilísticos para termos. Finalmente, trata-se de realizar extração dos candidatos a termos, a partir da utilização de corpora paralelos, respondendo assim aos objetivos específicos motivacionais à construção de um glossário terminológico.

4.1.2 Tipo do Corpus

O tipo de *corpus* a ser utilizado no trabalho precisa estar alinhado ao item: "objetivo de criação do corpus, que estará, consequentemente, alinhado com o objetivo da pesquisa". Como pode ser observada, a necessidade de textos paralelos para a realização do alinhamento e, posteriormente, para a extração terminológica, apresenta de antemão algumas pistas sobre a escolha do tipo de *corpus* que deve ser utilizado na pesquisa.

A presente investigação é então realizada sobre um *corpus* paralelo, com a finalidade de se buscar candidatos a termos, através de sua base de alinhamento. Como abordado na seção *Corpus*/*Corpora* Paralelo(s), esse *corpus* constitui-se de um conjunto de textos e de suas respectivas traduções. Para que a definição do tipo do *corpus* seja mais bem entendida, esta seção apresenta-se divida em subseções, a saber:

(i) domínio do *corpus*;
(ii) número de línguas no *corpus*; e
(iii) direcionalidade do *corpus*.

4.1.3 Domínio do Corpus

Proposto por Baker (1995, p. 229), o domínio é categorizado em apenas dois tipos: geral ou específico. Por domínio geral entende-se aqui a concentração de todas as áreas de conhecimento; enquanto para domínio específico a concentração de uma área do conhecimento particular. Dessa forma, como o estudo objetiva realizar a investigação no âmbito da área das ciências médicas, o domínio do *corpus* será específico. Embora, naturalmente essa grande área possa ser filtrada em seu próprio escopo, como por exemplo, pediatria, neurologia, cardiologia, etc., o trabalho investiga a área integral das ciências médicas, como já afirmado, em razão da ausência de limites para o processamento ofertado pelo desempenho dos suportes informáticos atuais. Não seria pertinente recusar a gama apurada de memória e capacidades de processamento das atuais gerações de máquinas e programas de rodagem.

4.1.4 Número de Línguas

Na utilização de um *corpus* paralelo, pressupõe-se que há, no mínimo, duas línguas envolvidas. Baker (Ibid.) propõe a classificação quanto ao número de línguas envolvidas em monolíngue, bilíngue, e multilíngue. Por *corpus* paralelo bilíngue entende-se o envolvimento de duas línguas. Por *corpus* paralelo multilíngue entende-se o envolvimento de três ou mais línguas, ligeiramente diferente do que se mencionou acima.

O *corpus* usado no presente trabalho é constituído de duas línguas em sua base, o inglês e o português, tratando-se assim de um *corpus* paralelo bilíngue inglês/português.

4.1.5 Direcionalidade

Por direcionalidade entende-se a relação de tradução entre as línguas envolvidas no *corpus*, de acordo com Zanettin (2000). Essa relação de direcionalidade pode ser classificada em unidirecional, bidirecional e multidirecional. Por unidirecional entende-se a relação de tradução de L1 e respectiva L2. Por bidirecional entende-se a relação de tradução L1 e respectiva L2, como também L2 e respectivas traduções para L1. Por multidirecional entende-se a relação de tradução Ln para Ln. Como a relação de tradução que constitui o *corpus* é da ordem L1 para L2 somente, quanto ao quesito direcionalidade o processo será classificado como unidirecional.

4.1.6 Classificação do Corpus

Com base nos dados apresentados nas seções anteriores, a classificação geral do *corpus* configura-se da seguinte forma: *corpus* paralelo bilíngue, de domínio específico, e unidirecional. Para clarificar, a tabela Configuração do *Corpus* abaixo representa sua configuração completa.

4.1.7 Direitos Autorais

No que diz respeitos aos direitos autorais envolvidos nos corpora selecionados, observamos que os textos disponíveis no sítio da *European Medicines Agency* (*EMA*) caracterizados como "*Public*" têm suas propriedades livres para utilização nas mais diversas situações, podendo ser "reproduzido e/ou distribuído, totalmente ou em parte, independente dos meios e/ou formatos utilizados, para fins comerciais e não comerciais".

4.2 CONSTRUÇÃO DO CORPUS

Uma vez que a configuração do *corpus* estiver definida na etapa denominada Projeto do *Corpus*, a fase de Construção do *Corpus* constituirá o momento em que ocorre o contato com o material textual a ser utilizado para o processamento e sua preparação. A etapa de construção do *corpus* é considerada como a mais longa e penosa pela maioria dos pesquisadores. Tal constatação deve-se ao fato de se trabalhar diretamente sobre a preparação do material a ser incluído no *corpus*. No entanto, nos últimos anos essa ideia tem mudado bastante, haja visto o desenvolvimento de novas ferramentas passíveis de auxiliar o investigador nessa atividade e a automatização da tarefa de construção dos corpora.

Neste processo, três etapas são realizadas, sendo elas:

(i) obtenção dos textos;
(ii) preparação dos textos; e
(iii) alinhamento dos textos.

Estas três etapas serão descritas com maiores detalhes nas seções seguintes.

4.2.1 Coleta dos Textos

A referida etapa consiste em coletar todo o material textual apto à inclusão no *corpus* em construção. Como o material a ser utilizado é aquele disponibilizado pela EMA (www.ema.europa.eu), foi possível coletar todo o seu conteúdo em formato eletrônico, necessitando apenas realizar a conversão necessária e a codificação dos arquivos para o formato adequado ao processamento do *corpus*. Em se tratando de arquivos em texto puro (*raw text*), foi necessário eliminar todo e qualquer tipo de formatação, bem como todos e quaisquer espaços desnecessários.

Para automatizar o processo de captura, fez-se uso de ferramentas avançadas de busca, fornecidas pelo mecanismo de pesquisa do Google, em conjunto com o complemento para o Navegador *Mozilla Firefox Download ThemAll!*. A utilização dessas ferramentas, em conjunto, permitiu aperfeiçoar a fase de coleta dos textos, tornando a operação praticamente automática, após a definição de alguns parâmetros de configuração, definindo o formato dos arquivos a serem descarregados, como por exemplo, o modelo PDF. A Figura 3, a seguir, mostra uma captura de tela do complemento *Download ThemAll*!

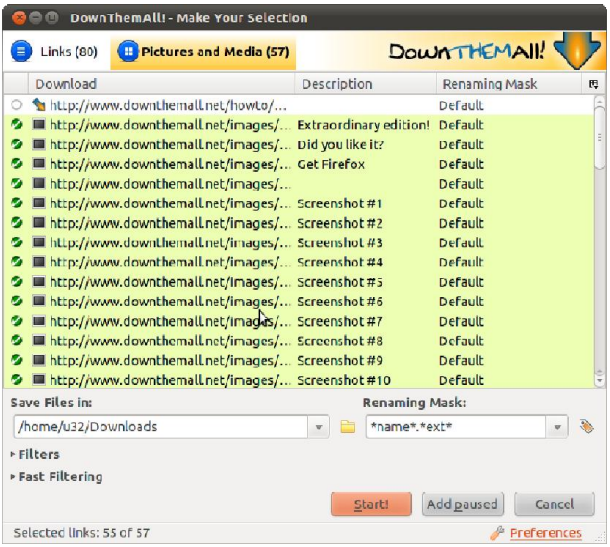

Figura 5: Captura de tela da ferramenta Download ThemAll!

Com os textos coletados, e considerando que a maioria das ferramentas utilizadas para processamento de *corpora* trabalha com arquivos em formato de texto simples (puro)[19], eliminando-se as marcações, converteu-se todo o material coletado para esse mesmo formato. A realização da conversão dos textos pode ser acompanhada na seção Conversão dos Textos.

4.2.2 Codificação dos Textos

Por codificação dos textos entende-se:

[19] Texto plano, ou texto puro, em informática, é um texto sem recursos de formatação tais como negrito, itálico e sublinhado. Mesmo sendo um formato simples, todos os compiladores das linguagens de programação e os programas que interpretam linguagens de marcação utilizam textos planos. O termo é também usado em criptografia para se referir a qualquer informação legível ou que possa ser usada diretamente por algum dispositivo eletrônico ou programa de computação. Ao ser submetido ao processo de criptografia o texto plano é convertido em algo não-inteligível, chamado de texto cifrado. (http://pt.wikipedia.org/wiki/Texto_plano).

> Codificação se refere ao processo de representar a informação de alguma forma. A linguagem humana é um sistema de codificação que nos representam informações em termos de sequências de unidades lexicais, e aqueles em termos de sequências de som ou gesto. A linguagem escrita é um sistema derivado da codificação pelo qual as sequências de unidades lexicais, sons ou gestos são representados em termos de símbolos gráficos que compõem algum sistema de escrita. (CONSTABLE, 2001, tradução nossa)[20].

A codificação dos textos variou de acordo com os estágios de processamento do *corpus*. Devido ao fato de serem utilizadas ferramentas diversas para seu processamento, a codificação dos textos do *corpus* foi divida em: (i) codificação para utilização com auxílio do programa *WordSmith Tools* e; (ii) codificação padrão para utilização, com o restante das ferramentas utilizadas, principalmente na etapa de processamento por meio do conjunto de ferramentas fornecido pelo *GIZA++*.

No desenvolvimento do processo de conversão dos arquivos, do formato PDF para o formato TXT, os arquivos gerados foram codificados no formato UTF-8. Esse formato permitiu o processamento dos alinhamentos estatísticos, utilizando-se do conjunto de ferramentas fornecido pelo *GIZA++*, bem como a manipulação dos textos para a montagem da interface de visualização do *corpus* paralelo. No entanto, para a realização do processamento do *corpus*, visando à obtenção das Listas de Palavras e também das Listas de Palavras-Chave, os arquivos contendo textos em língua portuguesa não proporcionaram os resultados desejados ao serem manipulados com auxílio das ferramentas *WordList* e *KeyWords* do programa *WordSmith Tools*. Os *erros* apresentados com base nas ferramentas do *WordSmith Tools* ocorreram devido ao não reconhecimento de caracteres acentuados, característicos da língua portuguesa. Para tornar possível o processamento adequado destes

[20] Encoding refers to the process of representing information in some form. Human language is an encoding system by which we represent information in terms of sequences of lexical units, and those in terms of sound or gesture sequences. Written language is a derivative system of encoding by which those sequences of lexical units, sounds or gestures are represented in terms of the graphical symbols that make up some writing system.

caracteres, percebeu-se a necessidade de conversão dos arquivos em formato UTF-8 para ANSI. Com os arquivos já codificados em formato ANSI, foi realizado o processamento adequado para a extração das Lista de Palavras e das Listas de Palavras-Chave.

As figuras a seguir exemplificam os dados gerados pelo programa *WordSmith Tools*. A Figura 4 expõe o processamento de Lista de Palavras com arquivos em língua portuguesa, em formato UTF-8. A Figura 5 demonstra o processamento de Listas de Palavras com arquivos em língua portuguesa, em formato ANSI.

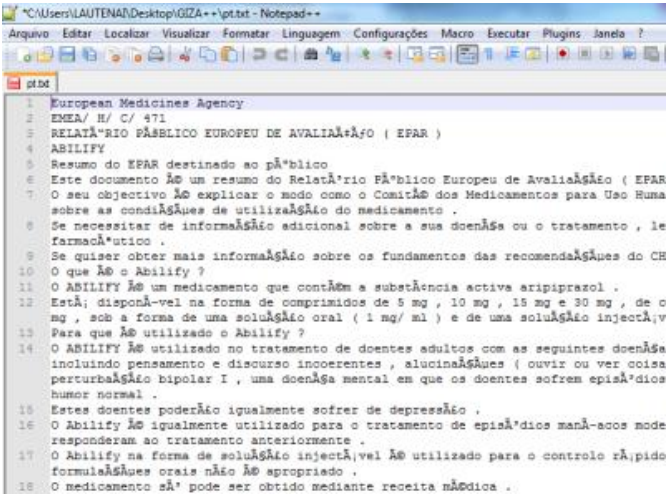

Figura 6: Formato do arquivo em ANSI.

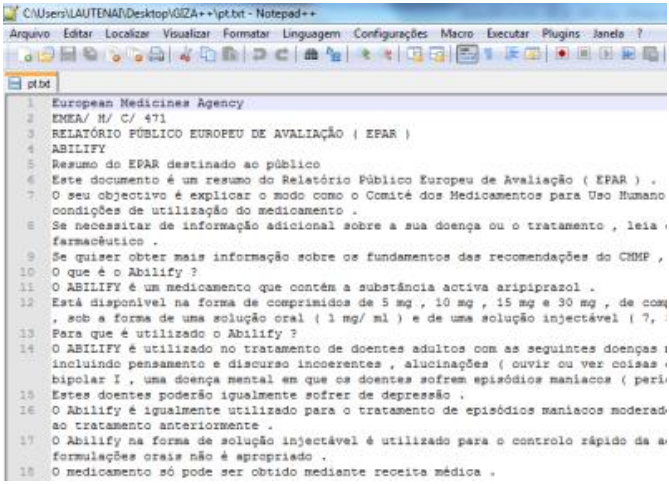

Figura 7: Formato do arquivo em UTF8.

4.2.3 Conversão dos Textos

Para a conversão dos textos coletados em formato PDF, utilizou-se da ferramenta *pdftotext* fornecida juntamente com a distribuição *Linux Ubuntu*. A sintaxe para uso é a seguinte: **$pdftotext<nome_do_arquivo>,** em que <**nome_do_arquivo**> refere-se à denominação da pasta no formato PDF, a ser convertido.

Como a necessidade de conversão era incontornável devido à enorme quantidade de arquivos em formato PDF, para que a tarefa de conversão dos arquivos não se tornasse cansativa, foi escrito um *script* com a finalidade de converter os arquivos PDF para TXT, em lotes. O referido *script* foi assim concebido e aplicado:

```
for f in *.pdf
do
pdftotext "$f"
done
```

Com a execução desse *script* no diretório dos arquivos em formato PDF, o material de cada pasta foi processado e convertido para o formato TXT, eliminando-se assim a necessidade de executar o *script* para cada arquivo tratado ou de abrir cada arquivo e salvar como TXT. Isso diminuiu consideravelmente o esforço e o tempo dispendido com a

tarefa de conversão dos arquivos para o formato de texto puro, geralmente conhecido como TXT.

4.2.4 Alinhamento do Corpus

Por alinhamento do *corpus* entende-se o processo de paralelizar os segmentos combinantes entre os arquivos constituintes do *corpus*.

O alinhamento do *corpus* foi divido em dois níveis, a saber:

(i) alinhamento ao nível de sentença; e

(ii) alinhamento ao nível de palavra.

O objetivo em realizar o alinhamento do *corpus* em dois níveis emergiu da necessidade de se ter uma interface de *corpus* paralelo que possibilite buscas ao nível de sentença, com a finalidade de apresentar buscas de contexto. O alinhamento ao nível de palavra é essencial para a realização dos procedimentos de extração de terminologia estatística, com base em alinhamento palavra-a-palavra.

Para a realização do alinhamento ao nível de sentença, foi utilizada a ferramenta de alinhamento *align*, desenvolvida por Church e Gale (1993). Com a utilização desse aparato, o procedimento de alinhamento dos textos foi realizado, justamente por comparar o material textual em língua inglesa com o material textual em língua portuguesa e, também, por gerar o arquivo com o segmento do texto e sua respectiva tradução.

4.2.5 Alinhador *Vanilla Aligner*

Para a realização do alinhamento ao nível de palavra e matriz de alinhamento, foi utilizado o conjunto de ferramentas fornecido pelo *GIZA++*, desenvolvido por Och (2003). Com a utilização dessa ferramenta, o procedimento de alinhamento foi realizado ao nível de palavra, possibilitando a geração de dados estatísticos para a probabilidade de cada unidade lexical e sua respectiva tradução. O método para utilizar do conjunto de ferramentas fornecido pelo *GIZA++* compreende uma série de etapas, dentre as quais a conversão dos arquivos em texto puro do *corpus* para o formato de leitura do conjunto de ferramentas fornecido pelo *GIZA++*, por meio da ferramenta

plain2snt.out; a geração de arquivos de vocabulários e classes, através da ferramenta *mkcls*; e, por fim, a realização do alinhamento, por meio da execução do conjunto de ferramentas fornecido pelo *GIZA++*. Essas etapas serão descritas em detalhes, a seguir.

4.2.6 Ferramenta plain2snt.out

A ferramenta *plain2snt.out*, parte integrante do *GIZA++*, tem por objetivo a conversão dos arquivos do *corpus*, do formato texto puro, TXT, para o formato aceito pelo do conjunto de ferramentas do *GIZA++*, SNT. Para a realização dessa conversão, a seguinte sintaxe foi utilizada:

$./plain2snt.ou arquivo1.txt arquivo2.txt

Na sintaxe de execução acima, tem-se como base a execução da ferramenta *plan2snt.out*, evocada através do comando *./plain2snt.out* e tendo como parâmetros o arquivo na língua A e o arquivo na língua B, *arquivo1.txt* e *arquivo2.txt*, respectivamente.

Os arquivos gerados pelo processamento dessa ferramenta são *arquivo1.vcb*, *arquivo2.vcb, arquivo1_arquivo2.snt* e *arquivo2_arquivo1.snt*. Em detalhes, esses arquivos representam cada qual:

arquivo1.vcb: arquivo contendo uma lista dos vocábulos extraídos da língua A no formato índice, vocábulo, número de ocorrências. A Figura 6, a seguir, reproduz um arquivo no formato *.vcb*, extraído da língua A.

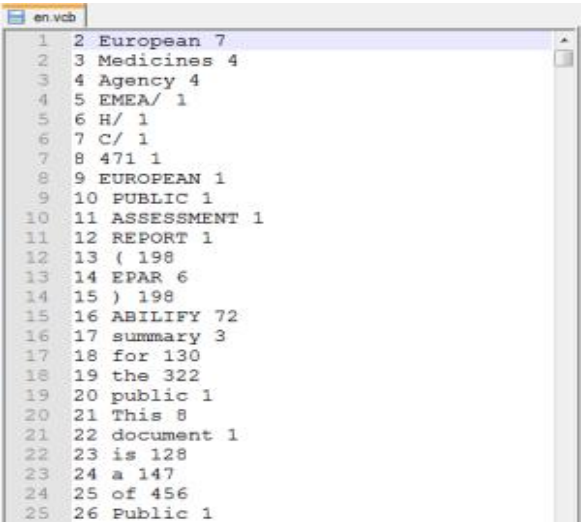

Figura 8: Formato do arquivo1.vcb.

arquivo2.vcb: arquivo contendo uma lista dos vocábulos extraídos da língua B no formato índice, vocábulo, número de ocorrências. A Figura 7, a seguir, representa um arquivo no formato *.vcb*, extraído da língua B.

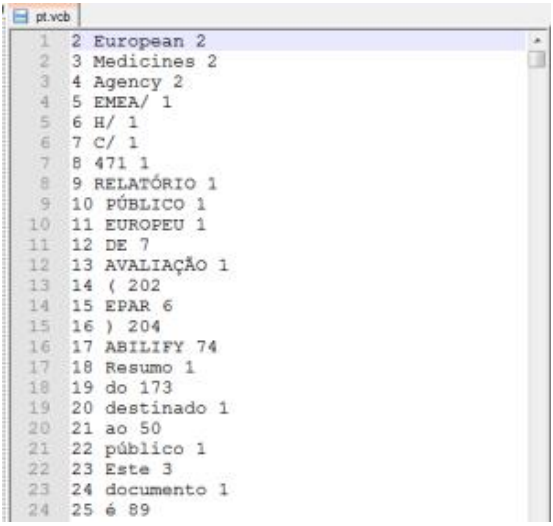

Figura 9: Formato do arquivo2.vcb.

arquivo1_arquivo2.snt: arquivo contendo uma matriz de alinhamento na direção língua A para língua B no formato números de ocorrências do alinhamento, índices dos vocábulos do arquivo1.vcb, índices dos vocábulos do arquivo2.vbc. A Figura 8, a seguir, expõe um arquivo no formato .snt extraído do alinhamento na direção da língua A para a língua B.

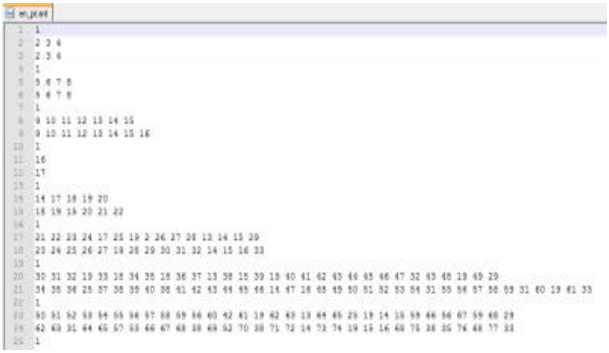

Figura 10: Formato do arquivo1_arquivo2.snt.

arquivo2_arquivo1.snt: arquivo contendo uma matriz de alinhamento na direção língua B para língua A no formato números de ocorrências do alinhamento, índices dos vocábulos do arquivo2.vcb, índices dos vocábulos do arquivo1.vbc. A Figura 9 a seguir expõe um arquivo no formato .snt extraído do alinhamento na direção da língua B para a língua A.

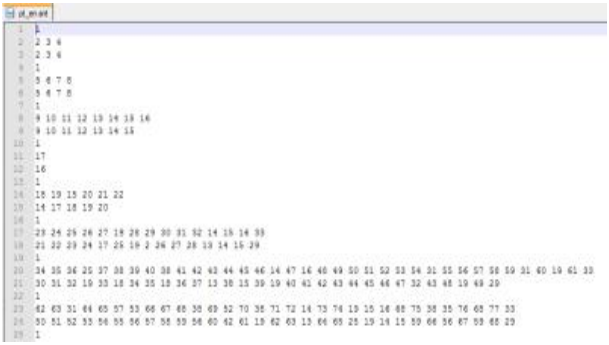

Figura 11: Formato do arquivo2_arquivo1.snt.

4.2.7 Ferramenta mkcls

A ferramenta *mkcls*, parte integrante do conjunto de ferramentas fornecido pelo *GIZA++*, tem por objetivo a geração de classes e categorias de palavras, a partir do *corpus* processado. Para a realização dessa operação, a seguinte sintaxe foi utilizada:

$mkcls -parquivo1.txt –Varquivo1.vcb.classes

$mkcls –parquivo2.txt –Varquivo2.vcb.classes

Na sintaxe da operação acima, tem-se como base a execução da ferramenta *mkcls*, chamada através do comando *./mkcls*, tendo como parâmetros o arquivo na língua A, no parâmetro –p, e o arquivo de classes na língua A, a ser gerado através do parâmetro –V, adicionando à extensão *.classes*, *-parquivo1.txt* e *–Varquivo1.txt*, respectivamente. O mesmo procedimento é realizado para a língua B.

Os arquivos gerados pelo processamento dessa ferramenta são *arquivo1.vcb.classes* e *arquivo2.vcb.classes*

4.2.8 Conjunto de Ferramentas *GIZA++*

A ferramenta *GIZA++* tem por objetivo realizar o alinhamento, com base estatística ao nível de palavra entre os textos constituintes do *corpus*. Para a realização dessa operação, a seguinte sintaxe é empregada:

$./GIZA++ -S arquivo.vcb -T arquivo2.vcb –C arquivo1_arquivo2.snt

Na sequência integrada de execução, acima reproduzida, tem-se como base a rodagem da ferramenta *GIZA++*, chamada através do comando *./GIZA++* , tendo como parâmetros o arquivo de vocabulário na língua A, no parâmetro –S (*source*), o arquivo de vocabulário na língua B, no parâmetro –T (*target*) e o arquivo de alinhamento entre classes na língua A, que será gerado através do parâmetro –V, adicionando à extensão *.classes, -parquivo1.txt e –Varquivo1.txt*, respectivamente. O mesmo procedimento é realizado para a língua B.

Os arquivos gerados pelo processamento dessa ferramenta são *arquivo1.vcb.classes* e *arquivo2.vcb.classes*.

4.3 PROCESSAMENTO DO CORPUS

A etapa de processamento do *corpus* assume a característica de produzir os dados a partir das ferramentas utilizadas, vislumbrando posterior análise. Nessa seção serão apresentadas as ferramentas utilizadas, o modo de processamento e os dados obtidos por meio desse processamento.

4.3.1 Ferramentas de Processamento do Corpus

Neste trabalho, de caráter investigativo, diferentes ferramentas com a finalidade de processar os materiais linguísticos foram utilizadas e serão apresentadas nas seções subsequentes com a finalidade de compor e demonstrar a proposta de ordem sequencial e criação de sistemas informáticos para extração terminológica bilíngue em corpora paralelos – inglês/português – com vistas à tradução de textos das ciências médicas, tal como pontuado no título da presente investigação. Posteriormente, como visto no modelo adotado para extração terminológica, serão apresentadas as ferramentas comporão a base para a proposta de extração bilíngue automatizada.

4.3.1.1 Emprego do WordSmith Tools 5

O nome *WordSmith Tools* na verdade não se refere a um único programa computacional, mas a um conjunto de ferramentas desenvolvidas para a análise textual, muito utilizado para pesquisas na área de linguística de *corpus*, linguística computacional e estudos da tradução baseados em *corpora*. A primeira versão do programa foi desenvolvida em 1997, por Mike Scott, na *Oxford University*. Em 2009, foi lançada a versão 5.0, versão mais atual, cuja forma foi utilizada para o presente trabalho.

A ferramenta *WordSmith Tools* oferece três recursos principais, são eles:

(i) *Concord,*
(ii) *WordList*, e
(iii) *KeyWords*.

Concord: ferramenta que tem por objetivo realizar linhas de concordâncias por meio da inserção de uma palavra de busca e apresentá-las de maneira horizontalmente paralela, juntamente com o contexto para ambos os lados da palavra buscada. O *Concord* oferece, por padrão, o formato de apresentação de linhas de concordâncias conhecido como KWIC (*Key-Word In Context* - Palavra-Chave no Contexto). O formato de apresentação das linhas de concordância em KWIC dispõe a palavra buscada centralizada na linha de concordância, facilitando sua percepção e manipulação. A Figura 10, a seguir, expõe uma captura de tela da ferramenta *Concord*, apresentando as linhas de concordância no formato KWIC.

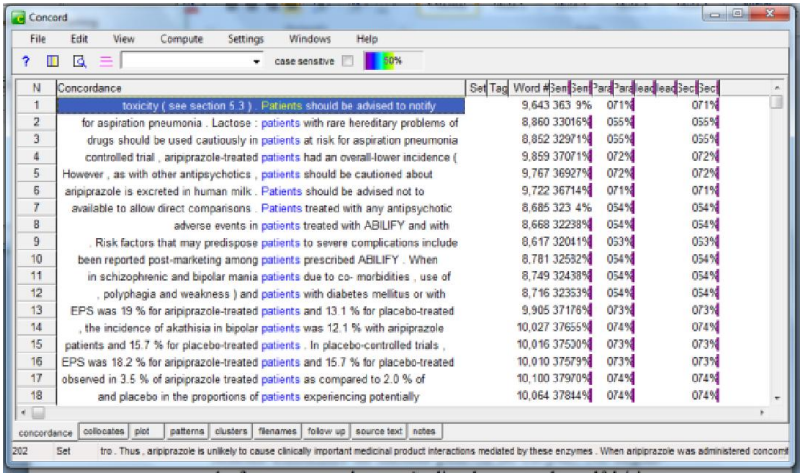

Figura 12: Linhas de concordâncias do Concord.

WordList: ferramenta que tem por objetivo realizar a catalogação de todas as palavras constituintes do *corpus*. A *WordList* realiza a contagem de palavras no *corpus*, identifica as palavras distintas e lista-as. A apresentação das listas de palavras é, geralmente, fornecida em dois formatos distintos: por frequência e/ou por ordem alfabética. No primeiro formato, as palavras são classificadas em ordem de frequência, ou seja, de acordo com o número de ocorrências existentes da mesma palavra no *corpus*. No segundo formato, as palavras são listadas em ordem alfabética, ou seja, são listadas em ordem crescente ou decrescente juntamente com os dados de ocorrência para cada uma delas. A Figura 11 e Figura 12, a seguir, mostram uma captura de tela da

ferramenta *WordList* apresentando as listas de palavras classificadas por frequência e por ordem alfabética, respectivamente.

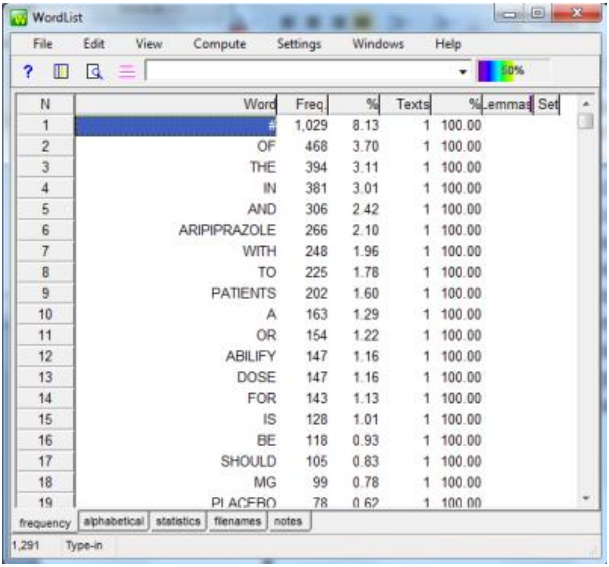

Figura 13: WordList por ordem de frequência.

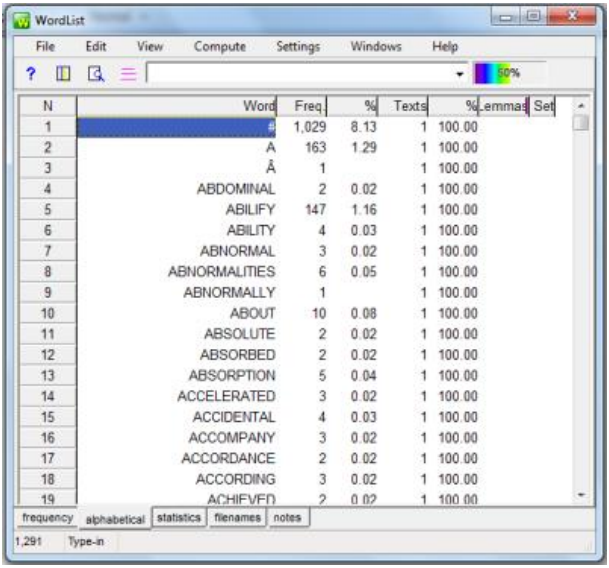

Figura 14: WordList por ordem alfabética.

A ferramenta de *WordList* é base para o trabalho posterior que será realizado com o auxílio da ferramenta *KeyWords*.

Como se pode constatar na Figura 11, na lista de palavras gerada pela ferramenta *WordList* do *WordSmith Tools*, itens considerados comuns em textos de todas a áreas, neste caso não somente ligadas à área das ciências médicas, também são processadas pela programa. No entanto, como o objetivo desse estudo não está voltado ao uso dessas palavras comuns, manifestadas muitas vezes na forma e função de artigos, preposições, interjeições, etc. usa-se uma configuração na ferramenta *WordList* que remove todas essas palavras no processamento dos corpora. Esse recurso é chamado de *Stop Words*, em tradução do tipo *literal*, isto é: *palavras de parada*, e tem por finalidade remover, por meio de instruções indicadas pelo usuário, os elementos sem interesse para o estudo. Tal procedimento visa, sobretudo, otimizar e pontuar os resultados que se busca.

Esse recurso é largamente utilizado por buscadores *online* como, por exemplo, o *Google*[21], *Bing*[22], *Duck Duck Go*[23], *Wikipédia*[24]

[21] https://www.google.com.br/
[22] http://bing.com/
[23] https://duckduckgo.com/

etc. Essa prática faz com que a busca por palavras comuns na maioria dos sítios *online* não sejam indexados nos resultados da busca e que apenas palavras realmente relevantes e relacionadas àquelas que foram inseridas na caixa de endereço sejam recuperadas. Para uma melhor definição temos:

> Palavras que não aparecem no índice de uma base de dados específica porque são insignificantes (i.e., artigos, preposições) ou tão comuns que o resultado seria maior do que o sistema pode lidar com (como no caso do IUCAT onde termos tais como United Estados ou Departamento são palavras de parada na busca por palavra-chave). Palavras de Parada variam de sistema para sistema. Além disso, alguns sistemas simplesmente ignoram palavras de parada pois o uso de palavras de parada em outros sistemas iriam resultar na recuperação de nenhum resultado. (SCHRUZ, F. D., 2012)[25]

Um exemplo de uma sentença composta apenas por palavras de parada pode ser encontrado no excerto de William Shakespeare, Hamlet, em inglês "*to be or not to be*". O excerto traz uma estrutura bastante comum na língua inglesa em sua composição que em uma busca pode gerar uma quantidade imensa de resultados não relevantes. Como já informado, com base no objetivo de cada pesquisa define-se a importância ou não de se manter essas palavras no processamento dos corpora. Tendo em vista que o objetivo do processamento dos corpora nessa pesquisa é recuperar resultados relevantes à área das ciências medicas, faz-se importante o uso das palavras de parada com a finalidade de obter resultados mais claros e reduzir em um terço o tempo de processamento dos corpora.

Para esse trabalho, foi utilizada uma lista com cerca de 600 palavras de parada. Um exemplo do arquivo de Palavras de Parada

---

[24] http://wikipedia.org/

[25] Words that do not appear in the index in a particular database because they are either insignificant (i.e., articles, prepositions) or so common that the results would be higher than the system can handle (as in the case of IUCAT where terms such as United States or Department are stop words in keyword searching.) Stop words vary from system to system. Also, some systems will merely ignore stop words where use of stop words in other systems will result in retrieving zero hits.

utilizado no *WordSmith Tools* contendo os primeiros itens é mostrado na sequência. Note-se que as primeiras linhas contendo um ponto e vírgula (;) no início da linha são comentários para entendimento do usuário e não são lidas pela ferramenta de geração de listas de palavras, a *WordList*. Esse tipo de comentário em arquivos é comum no campo da documentação, tornando o processo de entendimento de qualquer procedimento realizado de fácil compreensão, sem afetar nem o código processado pela máquina e nem o resultado fornecido.

; A potential stop list.
; Put ; at left margin to ignore a line.
; Separate items with commas.
; Use a Windows text editor, eg. NOTEPAD.
; Use CAPITALS.
#
A
A'S
ABLE
ABOUT
ABOVE
ACCORDING
ACCORDINGLY
ACROSS
ACTUALLY
AFTER
AFTERWARDS
AGAIN
AGAINST
AIN'T
ALL
ALLOW
ALLOWS
ALMOST
ALONE
ALONG
ALREADY
ALSO
ALTHOUGH
ALWAYS
AM
AMONG
AMONGST
AN
AND
ANOTHER
ANY
ANYBODY
ANYHOW
ANYONE
ANYTHING
ANYWAY
ANYWAYS
ANYWHERE
APART
APPEAR
APPRECIATE
APPROPRIATE
ARE
AREN'T
AROUND
AS
ASIDE
ASK
ASKING
ASSOCIATED
AT
AVAILABLE
AWAY
AWFULLY

BE
BECAME
BECAUSE
BECOME
BECOMES
BECOMING
BEEN
BEFORE
BEFOREHAND
BEHIND
BEING
BELIEVE
BELOW
BESIDE
BESIDES
BEST
BETTER
BETWEEN
BEYOND
BOTH
BRIEF
BUT
BY
C'MON
C'S
CAME
CAN
CAN'T
CANNOT
CANT
CAUSE
CAUSES
CERTAIN
CERTAINLY
CHANGES
CLEARLY
CO
COM
COME
COMES
CONCERNING
CONSEQUENTLY
CONSIDER
CONSIDERING
CONTAIN
CONTAINING
CONTAINS
CORRESPONDING
COULD
COULDN'T
COURSE
CURRENTLY
DEFINITELY
DESCRIBED
DESPITE
DID
DIDN'T
DIFFERENT
DO
DOES
DOESN'T
DOING
DON'T
DONE
DOWN
DOWNWARDS
DURING
EACH
EDU
EG
EIGHT
EITHER
ELSE
ELSEWHERE
ENOUGH
ENTIRELY
ESPECIALLY
ET
ETC
EVEN
EVER
...

Uma lista completa contendo todas as Palavras de Parada usadas para esta pesquisa pode ser encontrada no *Project Humboldt Digital Library*. Essa lista foi compilada para uso no projeto *Project Humboldt Digital Library*[26] por Armand Brahaj.

KeyWords: ferramenta que tem por objetivo realizar a catalogação das palavras-chave constituintes do *corpus*. Para a realização do cálculo de palavras-chave observou-se o cálculo de qui-quadrado baseado na operação da correção de Yates referente ao qui-quadrado. Isso se deve ao fato de o *corpus* da pesquisa não representar a totalidade do material produzido sobre o tema abordado. A Figura 13, a seguir, mostra uma captura de tela da análise das palavras-chave pela ferramenta *KeyWords.*

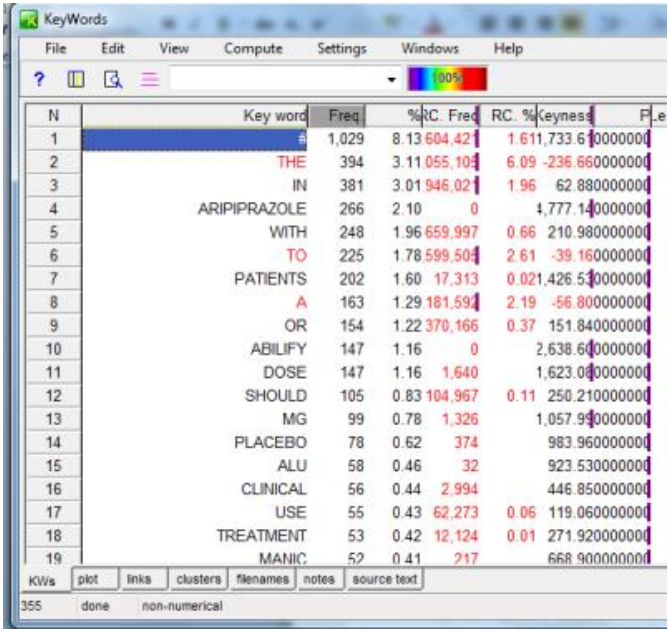

Figura 15: KeyWords.

Destaca-se na Tabela 2 a lista das 100 palavras com o valor de chavicidade mais elevado.

Tabela 2- 100 Palavras-Chave Corpus de Amostragem

N	KEY WORD	FREQ.	%	RC.	KEYNESS

[26] http://www.avhumboldt.net/

				FREQ.	
1	PATIENTS	11680	0,968189001	17313	64784,46875
2	DOSE	7002	0,580416083	1640	53638,52344
3	MG	4218	0,349642247	1326	31269,52734
4	INSULIN	3711	0,307615548	631	29263,56445
5	MEDICINAL	3290	0,272717625	263	27251,30273
6	HAEMOGLOBIN	3258	0,270065069	320	26689,16797
7	ML	3244	0,268904567	1709	22371,54102
8	INJECTION	3026	0,250833899	1008	22271,45117
9	TREATMENT	4234	0,35096851	12124	19062,41797
10	CLINICAL	3078	0,255144328	2994	18899
11	ALFA	2241	0,185762987	120	18887,625
12	DL	2170	0,179877579	73	18563,16992
13	ARANESP	2001	0,165868685	0	17709,10352
14	AUTHORISATION	2045	0,169515967	268	16446,11328
15	EU	1855	0,153766319	44	15999,54004
16	RENAL	1952	0,161806926	480	14870,81738
17	IU	1726	0,143073142	113	14428,29395
18	THERAPY	2242	0,185845867	1902	14171,32617
19	BLOOD	3176	0,263267845	9767	13922,1084
20	ADMINISTRATION	2843	0,235664502	6408	13902,43555
21	PRE	1760	0,145891502	441	13381,50586
22	ADMINISTERED	1853	0,153600529	971	12787,67383
23	DISORDERS	1794	0,148709849	834	12612,6543
24	TABLETS	1798	0,149041429	919	12457,34668
25	G	3346	0,277359635	15681	12296,27148
26	MMOL	1557	0,129064232	370	11902,91504
27	KG	1696	0,140586346	949	11579,3584
28	EPOETIN	1276	0,105771333	7	11205,25586
29	USE	5232	0,433695644	62273	11004,90527
30	SUBCUTANEOUS	1295	0,107346296	92	10785,41016
31	ROSIGLITAZONE	1202	0,099637263	0	10637,06543
32	ARIPIPRAZOLE	1199	0,099388584	0	10610,51367
33	EFFECTS	2674	0,221655607	10764	10514,66895
34	PRODUCT	2617	0,216930702	11039	10083,38086
35	STUDIES	2764	0,229115963	13572	9936,962891
36	TEL	1599	0,132545739	1608	9743,725586
37	ORAL	1711	0,141829744	2338	9682,833008
38	PLACEBO	1291	0,107014731	374	9659,881836
39	ANAEMIA	1283	0,106351584	363	9625,895508
40	MEDICINES	1346	0,111573838	611	9496,219727
41	SYMPTOMS	1754	0,145394146	3119	9229,97168
42	SYRINGE	1133	0,093917653	187	8953,832031
43	REACTIONS	1530	0,126826137	2023	8731,943359
44	LEAFLET	1351	0,111988299	1173	8497,609375

45	RECOMMENDED	1684	0,139591634	3719	8288,27832
46	EXCIPIENTS	919	0,076178573	1	8116,828125
47	ERYTHROPOIETIN	900	0,07460361	9	7863,536133
48	DOCTOR	2087	0,172997475	9352	7826,541016
49	TREATED	1905	0,157910973	6936	7811,899414
50	MARKETING	1687	0,13984032	5198	7387,702148
51	TABLET	1007	0,083473146	345	7383,82666
52	PLASMA	1157	0,095907077	970	7329,976074
53	PATIENT	1824	0,151196644	7282	7196,601074
54	OBSERVED	1635	0,135529891	5007	7176,525879
55	ACTRAPHANE	796	0,065982744	1	7028,578613
56	SOLUTION	1755	0,145477027	6801	7012,435059
57	ALLERGIC	921	0,076344356	247	6950,95752
58	SECTION	2423	0,200849488	18725	6839,140137
59	DARBEPOETIN	767	0,063578852	0	6787,271484
60	S	2622	0,217345178	23425	6759,251953
61	ABILIFY	752	0,062335458	0	6654,525879
62	FILLED	1547	0,12823531	5039	6631,589844
63	INTRAVENOUS	920	0,076261461	367	6611,508789
64	APPROXIMATELY	1318	0,10925284	2829	6546,330566
65	DOSES	986	0,081732392	681	6486,978027
66	PHARMACOKINETIC	731	0,060594708	5	6408,918457
67	L	1954	0,161972716	11553	6405,135254
68	RECEIVING	1331	0,110330448	3239	6342,960938
69	ADVERSE	1050	0,087037541	1181	6235,139648
70	RISK	1926	0,159651712	11759	6207,859863
71	PHARMACIST	774	0,064159103	110	6187,689453
72	IMPAIRMENT	875	0,072531283	396	6175,670898
73	BENEFIX	689	0,057113204	0	6096,996582
74	UNCOMMON	929	0,077007502	661	6078,057617
75	CHEMOTHERAPY	820	0,067972176	267	6050,774902
76	SODIUM	941	0,078002214	916	5775,208984
77	INCREASED	1901	0,157579392	12963	5769,179199
78	WEEKS	2007	0,166366041	15114	5752,430176
79	DESLORATADINE	648	0,053714596	0	5734,163574
80	LT	1216	0,100797758	3127	5685,976563
81	VIAL	722	0,05984867	137	5638,444336
82	DOSING	693	0,057444777	84	5602,109863
83	PRECAUTIONS	845	0,070044495	578	5569,182129
84	TRIALS	1079	0,089441434	2091	5533,089355
85	PAEDIATRIC	699	0,057942133	120	5505,952637
86	CONTAINS	1312	0,108755477	4570	5477,008301
87	DIABETES	832	0,068966888	652	5342,844238
88	PRODUCTS	1673	0,138679817	10587	5289,888184
89	AUC	601	0,049818631	7	5241,986816

90	PARTICULARS	803	0,066562995	611	5186,51709
91	IX	726	0,060180243	319	5146,230957
92	HYPOGLYCAEMIA	666	0,055206668	142	5145,638184
93	GLUCOSE	791	0,065568283	602	5108,786133
94	EMEA	577	0,0478292	0	5105,850098
95	RECOMBINANT	690	0,057196099	221	5101,696777
96	µG	573	0,04749763	0	5070,452148
97	LIVER	950	0,078748249	1632	5049,005859
98	DOSAGE	638	0,052885666	133	4939,810547
99	INHIBITORS	667	0,055289563	240	4859,918945
100	PREGNANCY	913	0,075681217	1562	4858,095215

Para observação inicial, foram utilizadas as primeiras 100 palavras-chave geradas pela ferramenta *KeyWords* do *WordSmith Tools*. Para a análise final, temos a lista atualizada com as 100 primeiras palavras-chave. Percebe-se que na lista apresentada na sequência, mesmo com o processamento do corpus completo, a ordem das palavras consideradas candidatos a termos não sofreram alterações significativas. Isso evidencia a relevância da primeira constatação realizada com o corpus de amostragem.

N	**KEY WORD**	**FREQ.**	**%**	**RC. FREQ.**	**KEYNESS**
1	PATIENTS	95987	0,770368695	17313	329277,9063
2	MG	69128	0,554804802	1326	291005,875
3	DOSE	58567	0,470044732	1640	242723,6563
4	TREATMENT	48536	0,389538318	12124	155436,8594
5	MEDICINAL	34303	0,275307685	263	147667,7656
6	ML	31752	0,254833966	1709	126389,875
7	INJECTION	29703	0,238389179	1008	121844,8516
8	INSULIN	28217	0,226462886	631	118023,8906
9	AUTHORISATION	21768	0,17470476	268	92776,09375
10	BLOOD	30031	0,241021618	9767	89865,64844
11	CLINICAL	24165	0,193942502	2994	87997,54688
12	EFFECTS	30089	0,241487116	10764	87592,38281
13	ADMINISTRATION	26818	0,215234846	6408	86724,49219
14	PRODUCT	28610	0,229617015	11039	81373,58594
15	DISORDERS	19609	0,157377139	834	79350,24219
16	EU	17685	0,141935572	44	77065,26563
17	DOCTOR	26348	0,211462751	9352	76875,15625
18	THERAPY	20157	0,161775261	1902	76021,29688
19	TABLETS	18927	0,15190357	919	75901,72656
20	TEL	19550	0,156903625	1608	74864,875
21	ORAL	19039	0,152802452	2338	69411,71875

22	RENAL	16513	0,132529393	480	68264,74219
23	MEDICINES	16678	0,133853644	611	68105,91406
24	LEAFLET	15958	0,128075093	1173	61806,27734
25	MARKETING	18239	0,146381855	5198	56525,98828
26	ADMINISTERED	14381	0,115418464	971	56144,93359
27	RECOMMENDED	16782	0,134688318	3719	55165,25391
28	REACTIONS	15203	0,12201564	2023	54779,85547
29	KG	13960	0,112039618	949	54466,96875
30	STUDIES	21755	0,174600422	13572	51689,70703
31	SYMPTOMS	14954	0,120017223	3119	49783,52734
32	SOLUTION	17131	0,137489304	6801	48272,17188
33	SECTION	22850	0,183388636	18725	47550,66406
34	VIAL	10845	0,087039374	137	46183,61719
35	DOSES	11410	0,091573931	681	45026,12109
36	TREATED	16085	0,129094362	6936	44101,5
37	PLACEBO	10760	0,086357184	374	44066,05469
38	TABLET	10701	0,08588367	345	44002,86719
39	PRE	10668	0,085618816	441	43241,41016
40	ADVERSE	11569	0,092850029	1181	43214,65234
41	PRODUCTS	17661	0,14174296	10587	42691,47656
42	RISK	17672	0,141831249	11759	40781,71875
43	HAEMOGLOBIN	9717	0,077986315	320	39911,05859
44	COATED	9787	0,078548118	415	39605,25781
45	PLASMA	10349	0,083058596	970	39054,23438
46	EMEA	8834	0,070899568	0	38792,42578
47	SEVERE	13171	0,105707295	4570	38675,65234
48	OBSERVED	13429	0,107777938	5007	38591,45313
49	SYRINGE	9169	0,0735882	187	38474,25391
50	INCREASED	17434	0,139921114	12963	38144,80469
51	ALFA	8817	0,070763133	120	37473,17578
52	SPECIAL	20771	0,166703075	21868	37300,87891
53	COMBINATION	12618	0,101269051	4395	37009,59766
54	PREGNANCY	10294	0,082617179	1562	36332,69922
55	IMPAIRMENT	8936	0,071718201	396	36056,30078
56	DAILY	14012	0,112456962	7519	35449,13281
57	PRECAUTIONS	8964	0,071942918	578	35138,16406
58	EXCIPIENTS	7998	0,064190038	1	35101,11328
59	PHARMACIST	8246	0,066180423	110	35064,76172
60	SUBCUTANEOUS	8164	0,065522313	92	34861,1875
61	HEPATIC	8218	0,065955698	304	33535,03906
62	SKIN	12683	0,101790726	6700	32284,65625
63	LIVER	9268	0,074382752	1632	31878,83203
64	DOSAGE	7492	0,060128998	133	31589,34961
65	ACTIVE	12735	0,102208063	7238	31477,32422
66	MEDICINE	10029	0,080490358	2728	31442,40625

67	SODIUM	8398	0,067400336	916	31105,96484
68	HOLDER	10084	0,080931775	2973	30974,13086
69	OLANZAPINE	6958	0,055843245	0	30553,48242
70	REPORTED	14619	0,117328599	11927	30487,15039
71	UNCOMMON	7938	0,063708492	661	30351,65625
72	DISEASE	13187	0,105835706	8869	30277,89258
73	BATCH	7958	0,063869007	794	29807,67578
74	INFUSION	7571	0,060763035	477	29737,26953
75	IRBESARTAN	6717	0,053909034	0	29495,10547
76	PARTICULARS	7600	0,060995784	611	29166,79492
77	DL	6768	0,05431835	73	28928,20703
78	HYPOGLYCAEMIA	6794	0,054527018	142	28481,30859
79	PACKAGE	11105	0,089126073	5859	28280,73242
80	NEEDLE	7995	0,064165957	1214	28213,875
81	RIBAVIRIN	6318	0,050706755	1	27723,60547
82	PHARMACEUTICAL	7098	0,056966849	494	27630,19336
83	TRIALS	8524	0,068411581	2091	27389,98242
84	APPROXIMATELY	8964	0,071942918	2829	27035,96875
85	COMMON	16373	0,131405786	19872	26684,33008
86	RARE	9933	0,079719879	4571	26620,66016
87	PAIN	11148	0,089471184	7040	26338,0293
88	EXPIRY	6562	0,052665044	343	26167,10352
89	EFFICACY	6740	0,054093629	513	26010,54492
90	PATIENT	11167	0,089623667	7282	26005,0957
91	WEEKS	14127	0,113379925	15114	25101,41211
92	INHIBITORS	6095	0,04891701	240	24778,25781
93	MMOL	6229	0,049992464	370	24588,39844
94	SUBSTANCE	7811	0,062689215	2158	24392,96484
95	ALLERGIC	5982	0,0480101	247	24247,09961
96	INTRAVENOUS	6087	0,048852805	367	23997,84766
97	BREAST	7186	0,057673115	1621	23526,66211
98	GLUCOSE	6144	0,049310274	602	23062,76953
99	BLISTER	5360	0,043018077	74	22770,45703
100	INFECTIONS	6109	0,049029373	705	22458,33594

4.3.1.2 Vanilla Aligner

Nessa etapa da pesquisa, a utilização da ferramenta *VanillaAligner* permite realizar o alinhamento do *corpus* entre a língua A e língua B com a finalidade de criar um ambiente de concordâncias paralelas para observação do alinhamento entre as sentenças. Visto que a ferramenta *WordSmith Tools* oferece somente o recurso de

concordância monolíngue, percebeu-se a necessidade de criar um ambiente em que fosse possível a realização de concordâncias bilíngues.

Através do alinhamento produzido pela ferramenta *VanillaAligner*, obteve-se o alinhamento da totalidade do *corpus* em formato paralelo. Desenvolveu-se então, em PERL, outra ferramenta que possibilitasse a leitura dos alinhamentos entre sentenças produzidas com a possibilidade de buscar por nódulos específicos no *corpus* paralelo, doravante alinhado. Utilizando uma linha de comando, possível ler o *corpus* alinhado e extrair linhas de concordâncias em formato bilíngue.

4.3.1.3 *GIZA++*

O *GIZA++* constitui-se de um conjunto de ferramentas com recursos que fornecem o alinhamento entre os itens lexicais presentes no *corpus* ao nível de palavra, baseado em elementos estatísticos. Nesse estágio de processamento do *corpus*, foi possível estabelecer a relação de tradução entre cada elemento lexical existente no *corpus*. O objetivo principal da utilização dessa ferramenta é a busca automatizada por candidatos à equivalente para um determinado item lexical, baseado na lista de palavras-chave extraída por meio da ferramenta *KeyWords*. A seguir, serão expostas tabelas exemplificativas, referentes aos dados obtidos através do uso do *GIZA++* e seus componentes, a saber: *plain2snt.out* e *mkcls*.

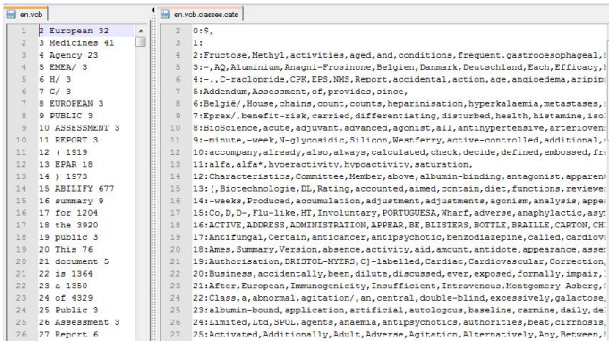

Figura 16: Vocábulos, Classes e Categorias - Corpus *em Inglês em GIZA++*

A Figura 14, acima, expõe a lista de vocábulos e a lista de classes e categorias para o processamento inicial do *corpus*, em língua inglesa. Esse processamento foi realizado por meio da ferramenta *mkcls*, apresentada na seção *mkcls*, durante a etapa de construção do *corpus*.

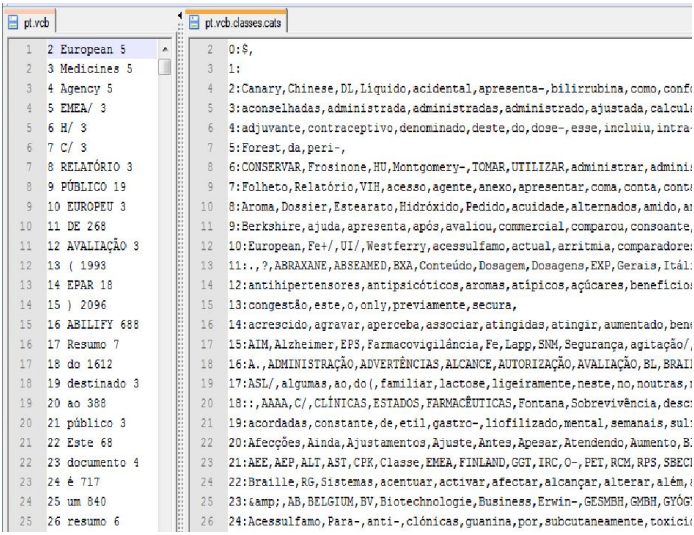

Figura 17: Vocábulos, Classes e Categorias - Corpus *em Português no GIZA++*

A Figura 15, acima, reproduz e evidencia a lista de vocábulos e a lista de classes e categorias para o processamento inicial do *corpus* em língua portuguesa. Esse processamento foi realizado por meio da ferramenta *mkcls*, apresentada na seção *mkcls*, durante a fase de construção do *corpus*.

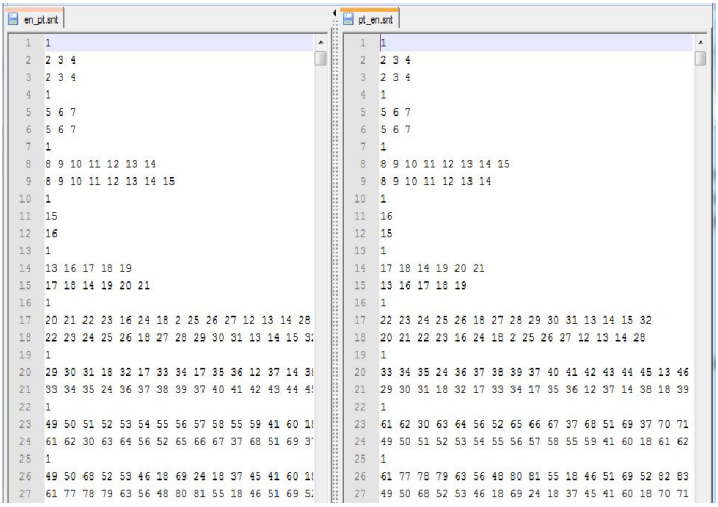

Figura 18: Alinhamento dos Vocábulos em Inglês e Português no GIZA++

A Figura 18 contempla uma matriz de alinhamento entre os arquivos de vocábulos do *corpus* em língua portuguesa e língua inglesa. O alinhamento foi realizado pelo índice de cada vocábulo no arquivo **.vcb**, produzido na etapa anterior. Como se observa, o alinhamento foi realizado ao nível de palavra, resultando em dois arquivos com alinhamento na direção inglês para português e, também, na direção do português para o inglês, evidenciados no lado esquerdo e no lado direito da figura, respectivamente.

Para melhor compreensão do referido procedimento, eis abaixo a visualização dos dados apresentados na Tabela 3.

EN.VCB	**PT.VCB**
2 European 32	2 European 5
3 Medicines 41	3 Medicines 5
4 Agency 23	4 Agency 5
5 EMEA/ 3	5 EMEA/ 3
6 H/ 3	6 H/ 3
7 C/ 3	7 C/ 3
8 EUROPEAN3	8 RELATÓRIO 3
9 PUBLIC 3	9 PÚBLICO 19
10 ASSESSMENT 3	10 EUROPEU 3
11 REPORT 3	11 DE 268
12 (1919	12 AVALIAÇÃO 3

13 EPAR 18	13 (1993
14) 1973	14 EPAR 18
	15) 2096

Tabela 3: Matriz de alinhamento 1.

Na tabela 3, são apresentados os arquivos de vocábulos no formato: índice de cada palavra no *corpus*, palavra no *corpus* e, então, a frequência da palavra no *corpus*. Esse formato se deve ao fato de que, a partir desse momento, o alinhamento será baseado somente no índice de cada palavra. Na Tabela 4 será apresentado o alinhamento, em ambas as direções, para as sentenças.

EN_PT.SNT	**PT_EN.SNT**
1	1
2 3 4	2 3 4
2 3 4	2 3 4
1	1
5 6 7	5 6 7
5 6 7	5 6 7
1	1
8 9 10 11 12 13 14	8 9 10 11 12 13 14 15
8 9 10 11 12 13 14 15	8 9 10 11 12 13 14

Tabela 4: Matriz de alinhamento 2.

Como observado na tabela 4, o alinhamento de cada sentença se conforma ao índice de cada palavra no *corpus*, como visto na tabela anterior. O exemplo a seguir mostra uma sentença no alinhamento ao nível de palavra, na direção do inglês para português e português para inglês, respectivamente.

Exemplo 1:

8 9 10 11 12 13 14
8 EUROPEAN| 9 PULIC| 10 ASSESSMENT | 11 REPORT | 12 (| 13 EPAR |)
8 9 10 11 12 13 14 15
8 RELATÓRIO | 9 PÚBLICO | 10 EUROPEU | 11 DE | 12 AVALIAÇÃO | 13 (| 14 EPAR | 15)

Exemplo 2:

8 9 10 11 12 13 14 15
8 RELATÓRIO | 9 PÚBLICO | 10 EUROPEU | 11 DE | 12 AVALIAÇÃO | 13 (| 14 EPAR | 15)
8 9 10 11 12 13 14

Os exemplos apresentados elucidam o modelo de alinhamento obtido a partir do índice de vocábulos inerente a cada palavra no *corpus*. Com estes arquivos processados, o próximo estágio visará à execução da ferramenta *GIZA++* para obtenção dos dados estatísticos entre os itens lexicais do *corpus*. Nesse estágio de processamento, um conjunto de arquivos é gerado de forma que sejam relacionados entre si, com a finalidade de apresentar os dados para possível análise por uma ferramenta de visualização de dados. A Figura 17, abaixo, expõe uma captura de tela com todos os arquivos gerados pelo processamento da ferramenta *GIZA++*.

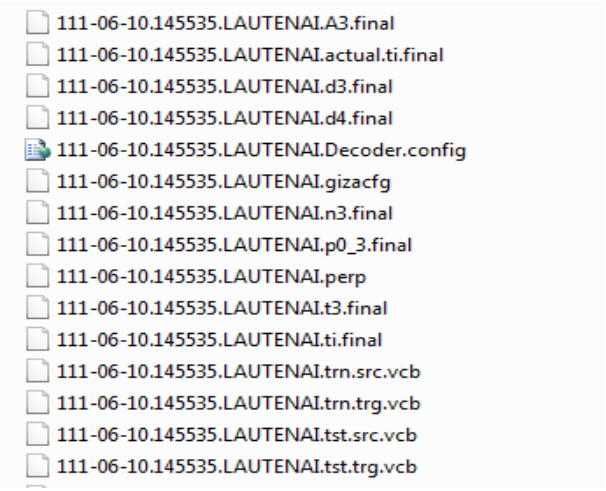

Figura 19: Arquivos produzidos pela GIZA++

Os arquivos processados pela ferramenta *GIZA++* serão posteriormente analisados pela ferramenta *Cairoize*, que os converterá para o formato XML com os dados relacionados, de modo que seja possível sua visualização na ferramenta *Cairo* e, então, a análise pelo pesquisador. Nesse estágio, o principal arquivo utilizado para análise é o arquivo com a extensão **.actual.ti.final**. Esse arquivo apresenta, de forma legível, a probabilidade de tradução para cada item lexical. A Figura 18, a seguir, expõe, a título ilustrativo, uma captura de tela do arquivo **actual.ti.final**, gerado pela ferramenta *GIZA++*.

111-06-10.145535.LAUTENAI.actual.ti.final

```
358  recebiam receiving 1
359  pescoço receiving 0.558005
360  eventualidade event 1
361  médicos attention 0.624966
362  locais reuptake 0.192513
363  recaptação reuptake 1
364  grande dopamine 0.225673
365  blister blister 0.925325
366  do cell 0.0184542
367  dopaminérgicos dopamine 1
368  Na one-half 0.0623854
369  Observou- Anaemia 0.414241
370  anemia Anaemia 0.002704
371  da mortality 0.00309528
372  nas minimisation 0.00853429
373  arritmias arrhythmias 1
374  as marketed 0.0397079
375  PET PET-bottles 0.142857
376  Desconhece- The 0.0640819
377  recebido treatments 0.10696
378  sangue The 0.00538974
379  tratamentos treatments 0.330262
380  uma guidance 0.00239964
381  locais sites 0.0320856
382  metade one-half 0.986301
383  Deve The 0.472493
384  da adjustment 0.0182093
385  dose adjustment 0.0394902
386  agitada shaken 1
387  España España 1
388  mortalidade mortality 0.916337
389  é Erypo 0.000560115
390  do body 0.00205003
391  inflexibilidade inflexibility 1
```

Figura 20: Arquivo actual.ti.final em GIZA++

O arquivo apresentado na Figura 20, acima, é disposto no seguinte formato: item lexical em língua A, item lexical em língua B e nível de probabilidade de tradução entre eles, no *corpus*. Para nível de probabilidade 1, a possível tradução é considerada conforme. Outros níveis apresentados, considerando-se o valor maior aproximado ao valor de 1.0 serão tomados como melhores candidatos de tradução para os itens lexicais. No decorrer desta investigação, será desenvolvida uma ferramenta que agrupe e ordene todos os itens lexicais, por níveis de probabilidade de tradução, de modo a otimizar o trabalho de extração. Neste estágio do trabalho, a ferramenta *Cairo* será utilizada para a visualização das sentenças alinhadas ao nível de palavra. Para empregar tempestivamente essa ferramenta, primeiro será necessária a conversão dos dados produzidos pela ferramenta *GIZA++*, por meio da ferramenta *Cairoize*, cuja descrição segue abaixo.

4.3.1.4 Cairoize

A ferramenta *Cairoize* tem por objetivo realizar a conversão dos arquivos produzidos pela *GIZA++*, de forma a possibilitar a leitura

visual e legível ao humano através da ferramenta *Cairo*. Por fazer uso dos dados presentes nos arquivos relacionados, a *Cairoize* reúne, em apenas um arquivo XML, todas as informações da relação de sentenças e palavras do *corpus*.

Nesta etapa, com a finalidade tão somente de observar os resultados obtidos, aplicou-se a ferramenta *Cairoize*. Acreditou-se ser importante poder visualizar a disposição dos dados processados. Intenciona-se propor, futura e brevemente, uma ferramenta de melhor integração, elaborada especificamente para responder a tal finalidade, e com opções de distribuição do material por meio eletrônico.

5 DADOS: ANÁLISE

Nessa seção são apresentadas análises realizadas a partir do processamento dos corpora e da proposta de ordem sequencial para a criação de glossários bilíngues. Nas subseções que seguem, torna-se possível verificarmos os dados gerados por meio da metodologia propiciada pela Linguística de Corpus e suas ferramentas. Como visto, o emprego sistemático foi indispensável à testagem e geração dos dados com a finalidade de validar a proposta de ordem sequencial desenvolvida nesta investigação, respondendo assim aos objetivos e metas fixadas para a condução do trabalho.

5.1 CANDIDATOS A TERMOS

Na etapa embrionária da proposta sequencial utilizou-se, incialmente, os dados oferecidos pela ferramenta de geração palavras-chave. Seus resultados propiciaram pistas e rastros bastante seguros com vistas à determinação dos candidatos a termos e, a partir de então, sua eventual na recuperação nos procedimentos de busca por correspondentes tradutórios.

Na tabela abaixo são apresentados palavras-chave por ordem de frequência, seguidas das indicações correspondentes, representados em quantitativamente em números e em porcentagem, respectivamente. Em seguida, na quinta coluna, expõe-se a Relação de Frequência das palavras no Corpus de Referência, seguida do Índice de Chavicidade.

N	KEY WORD	FREQ.	%	RC. FREQ.	KEYNESS
1	PATIENTS	95987	0,770368695	17313	329277,9063
2	MG	69128	0,554804802	1326	291005,875
3	DOSE	58567	0,470044732	1640	242723,6563
4	TREATMENT	48536	0,389538318	12124	155436,8594
5	MEDICINAL	34303	0,275307685	263	147667,7656
6	ML	31752	0,254833966	1709	126389,875
7	INJECTION	29703	0,238389179	1008	121844,8516
8	INSULIN	28217	0,226462886	631	118023,8906
9	AUTHORISATION	21768	0,17470476	268	92776,09375
10	BLOOD	30031	0,241021618	9767	89865,64844
11	CLINICAL	24165	0,193942502	2994	87997,54688
12	EFFECTS	30089	0,241487116	10764	87592,38281
13	ADMINISTRATION	26818	0,215234846	6408	86724,49219

14	PRODUCT	28610	0,229617015	11039	81373,58594
15	DISORDERS	19609	0,157377139	834	79350,24219
16	EU	17685	0,141935572	44	77065,26563
17	DOCTOR	26348	0,211462751	9352	76875,15625
18	THERAPY	20157	0,161775261	1902	76021,29688
19	TABLETS	18927	0,15190357	919	75901,72656
20	TEL	19550	0,156903625	1608	74864,875
21	ORAL	19039	0,152802452	2338	69411,71875
22	RENAL	16513	0,132529393	480	68264,74219
23	MEDICINES	16678	0,133853644	611	68105,91406
24	LEAFLET	15958	0,128075093	1173	61806,27734
25	MARKETING	18239	0,146381855	5198	56525,98828
26	ADMINISTERED	14381	0,115418464	971	56144,93359
27	RECOMMENDED	16782	0,134688318	3719	55165,25391
28	REACTIONS	15203	0,12201564	2023	54779,85547
29	KG	13960	0,112039618	949	54466,96875
30	STUDIES	21755	0,174600422	13572	51689,70703
31	SYMPTOMS	14954	0,120017223	3119	49783,52734
32	SOLUTION	17131	0,137489304	6801	48272,17188
33	SECTION	22850	0,183388636	18725	47550,66406
34	VIAL	10845	0,087039374	137	46183,61719
35	DOSES	11410	0,091573931	681	45026,12109
36	TREATED	16085	0,129094362	6936	44101,5
37	PLACEBO	10760	0,086357184	374	44066,05469
38	TABLET	10701	0,08588367	345	44002,86719
39	PRE	10668	0,085618816	441	43241,41016
40	ADVERSE	11569	0,092850029	1181	43214,65234
41	PRODUCTS	17661	0,14174296	10587	42691,47656
42	RISK	17672	0,141831249	11759	40781,71875
43	HAEMOGLOBIN	9717	0,077986315	320	39911,05859
44	COATED	9787	0,078548118	415	39605,25781
45	PLASMA	10349	0,083058596	970	39054,23438
46	EMEA	8834	0,070899568	0	38792,42578
47	SEVERE	13171	0,105707295	4570	38675,65234
48	OBSERVED	13429	0,107777938	5007	38591,45313
49	SYRINGE	9169	0,0735882	187	38474,25391
50	INCREASED	17434	0,139921114	12963	38144,80469
51	ALFA	8817	0,070763133	120	37473,17578
52	SPECIAL	20771	0,166703075	21868	37300,87891
53	COMBINATION	12618	0,101269051	4395	37009,59766
54	PREGNANCY	10294	0,082617179	1562	36332,69922
55	IMPAIRMENT	8936	0,071718201	396	36056,30078
56	DAILY	14012	0,112456962	7519	35449,13281
57	PRECAUTIONS	8964	0,071942918	578	35138,16406
58	EXCIPIENTS	7998	0,064190038	1	35101,11328

59	PHARMACIST	8246	0,066180423	110	35064,76172
60	SUBCUTANEOUS	8164	0,065522313	92	34861,1875
61	HEPATIC	8218	0,065955698	304	33535,03906
62	SKIN	12683	0,101790726	6700	32284,65625
63	LIVER	9268	0,074382752	1632	31878,83203
64	DOSAGE	7492	0,060128998	133	31589,34961
65	ACTIVE	12735	0,102208063	7238	31477,32422
66	MEDICINE	10029	0,080490358	2728	31442,40625
67	SODIUM	8398	0,067400336	916	31105,96484
68	HOLDER	10084	0,080931775	2973	30974,13086
69	OLANZAPINE	6958	0,055843245	0	30553,48242
70	REPORTED	14619	0,117328599	11927	30487,15039
71	UNCOMMON	7938	0,063708492	661	30351,65625
72	DISEASE	13187	0,105835706	8869	30277,89258
73	BATCH	7958	0,063869007	794	29807,67578
74	INFUSION	7571	0,060763035	477	29737,26953
75	IRBESARTAN	6717	0,053909034	0	29495,10547
76	PARTICULARS	7600	0,060995784	611	29166,79492
77	DL	6768	0,05431835	73	28928,20703
78	HYPOGLYCAEMIA	6794	0,054527018	142	28481,30859
79	PACKAGE	11105	0,089126073	5859	28280,73242
80	NEEDLE	7995	0,064165957	1214	28213,875
81	RIBAVIRIN	6318	0,050706755	1	27723,60547
82	PHARMACEUTICAL	7098	0,056966849	494	27630,19336
83	TRIALS	8524	0,068411581	2091	27389,98242
84	APPROXIMATELY	8964	0,071942918	2829	27035,96875
85	COMMON	16373	0,131405786	19872	26684,33008
86	RARE	9933	0,079719879	4571	26620,66016
87	PAIN	11148	0,089471184	7040	26338,0293
88	EXPIRY	6562	0,052665044	343	26167,10352
89	EFFICACY	6740	0,054093629	513	26010,54492
90	PATIENT	11167	0,089623667	7282	26005,0957
91	WEEKS	14127	0,113379925	15114	25101,41211
92	INHIBITORS	6095	0,04891701	240	24778,25781
93	MMOL	6229	0,049992464	370	24588,39844
94	SUBSTANCE	7811	0,062689215	2158	24392,96484
95	ALLERGIC	5982	0,0480101	247	24247,09961
96	INTRAVENOUS	6087	0,048852805	367	23997,84766
97	BREAST	7186	0,057673115	1621	23526,66211
98	GLUCOSE	6144	0,049310274	602	23062,76953
99	BLISTER	5360	0,043018077	74	22770,45703
100	INFECTIONS	6109	0,049029373	705	22458,33594

Os candidatos a termos obtidos por meio do cálculo de qui-quadro representam os itens lexicais mais relevantes (índice de chavicidade elevada) no campo disciplinar das ciências medicais. Para a análise completa dos corpora, serão empregados os cálculos de probabilidade discutidos neste trabalho por meio de uma ferramenta chamada *WordStats*[27], desenvolvida com a finalidade de integrar a suíte de ferramentas para a ordem sequencial proposta nesta pesquisa.

Após a geração das palavras-chave baseada no corpus completo, apresentadas na tabela acima, os resultados permitem listas os dados como integrantes do pretendido glossário. Na próxima etapa de análise, buscaremos por meio do conjunto de ferramentas do *GIZA++* correspondentes dos candidatos a termos de forma automática.

Usaremos, para efeito de pesquisa apenas os primeiros 10 candidatos a termos com a finalidade de validar a proposta de ordem sequencial, após o processamento do corpus completo. Com efeito, no escopo deste texto, trata-se, naturalmente, de demonstrar por meio de exemplos e referências. Obviamente, não seria sensato, tampouco lógico, buscar a exaustividade na exposição de dados.

N	KEY WORD	FREQ.	%	RC. FREQ.	KEYNESS
1	PATIENTS	95987	0,770368695	17313	329277,9063
2	MG	69128	0,554804802	1326	291005,875
3	DOSE	58567	0,470044732	1640	242723,6563
4	TREATMENT	48536	0,389538318	12124	155436,8594
5	MEDICINAL	34303	0,275307685	263	147667,7656
6	ML	31752	0,254833966	1709	126389,875
7	INJECTION	29703	0,238389179	1008	121844,8516
8	INSULIN	28217	0,226462886	631	118023,8906
9	AUTHORISATION	21768	0,17470476	268	92776,09375
10	BLOOD	30031	0,241021618	9767	89865,64844

5.2 EXTRAÇÃO TERMINOLÓGICA BILÍNGUE

Nesta seção serão apresentadas análises com base na extração terminológica recuperada por meio do alinhamento automático ao nível de palavra. A seguir, serão analisados os correspondentes tradutórios obtidos por meio do processamento realizado para a detecção dos candidatos a termos no campo disciplinar das ciências medicais. A

[27] Disponível em maio de 2012 em: <http://lautenai.pos.ufsc.br/labs/wordstats/>

análise tomará como base o confrontamento do correspondente tradutório, obtido por meio do processamento automático, com dicionários monolíngues, bilíngues e ferramentas de apoio à tradução.

1	PATIENTS	82959	0,822645962	17313	307451,9063

O item lexical *patient* apresenta o maior índice de chavicidade positiva comparado aos demais termos do presente corpus. No material de estudo o item lexical *patient* ocorre 65646 vezes a mais do que no corpus de referência. Este índice de frequência, relativamente alto no corpus de estudo, pode significar que o termo é um provável candidato à palavra-chave.

Ao buscar o termo com base no alinhamento ao nível de palavra produzido pela ferramenta *GIZA* ++, percebe-se que o correspondente tradutório em língua portuguesa encontrado no corpus foi o item lexical *doente,* como pode ser observado no exemplo apresentado abaixo:

CANDIDATO	CORRESPONDENTE	PROBABILIDADE
Patients	doentes	0.0270459
Patients	Doentes	0.592483
patients	doentes	0.881883

Em seguida, abaixo, segue uma breve amostra de linhas de concordância paralela do termo *patient* e suas respectivas traduções em português.

TEXTO 1	TEXTO 2
Abilify is used to treat adults with the following mental illnesses : schizophrenia , a mental illness with a number of symptoms , including disorganised thinking and speech , hallucinations (hearing or seeing things that are not there) , suspiciousness and delusions (mistaken beliefs) ; bipolar I disorder , a mental illness in which **patients** have manic episodes (periods of abnormally high mood) , alternating with periods of normal	O ABILIFY é utilizado no tratamento de doentes adultos com as seguintes doenças mentais : • esquizofrenia , uma doença mental com numerosos sintomas , incluindo pensamento e discurso incoerentes , alucinações (ouvir ou ver coisas que não existem) , desconfiança e ilusões (juízos errados) ; • perturbação bipolar I , uma doença mental em que os **doentes** sofrem episódios maníacos (períodos de humor muito elevado) , alternados de

mood .	períodos de humor normal .
Abilify is used to treat moderate to severe manic episodes and to prevent manic episodes in **patients** who have responded to the medicine in the past .	O Abilify é igualmente utilizado para o tratamento de episódios maníacos moderados a graves e para a prevenção de episódios maníacos em **doentes** que responderam ao tratamento anteriormente .
The maintenance dose is 15 mg once a day , but higher doses may benefit some **patients** .	A dose de manutenção é de 15 mg uma vez por dia , mas alguns **doentes** podem beneficiar de doses superiores .
Some **patients** may benefit from a higher dose .	Alguns **doentes** podem beneficiar de uma dose mais elevada .
For both illnesses , the oral solution or orodispersible tablets can be used in **patients** who have difficulty swallowing tablets .	Em ambas as doenças , nos **doentes** que têm dificuldade em engolir comprimidos pode ser utilizada a solução oral ou os comprimidos orodispersíveis .
The daily dose of Abilify should not exceed 30 mg , but this dose should be used with caution in **patients** who have severe problems with their liver .	A dose diária de Abilify não deve exceder 30 mg . Esta dose deve ser utilizada com precaução em **doentes** com problemas hepáticos graves .
The dose of Abilify should be adjusted in **patients** who are taking other medicines that are broken down in the same way as Abilify .	A dose deve ser ajustada para os **doentes** que estejam a tomar outros medicamentos metabolizados da mesma forma que o Abilify .
For the treatment of schizophrenia , there were three main short-term studies of Abilify tablets lasting four to six weeks , which involved 1,203 **patients** and compared Abilify with placebo (a dummy treatment) .	No tratamento da esquizofrenia , foram efectuados três estudos principais de curta duração (quatro a seis 6 semanas) com Abilify na forma de comprimidos , que incluíram 1203 **doentes** e que compararam a eficácia do Abilify com a de um placebo (tratamento simulado) .
The effectiveness of the solution for injection was compared with	A eficácia da solução injectável foi comparada com placebo em

placebo over a period of two hours in two studies involving 805 **patients** with schizophrenia or related conditions who were experiencing symptoms of agitation .	dois estudos durante um período de duas horas que incluíram 805 **doentes** com esquizofrenia ou doenças relacionadas , que sofriam de sintomas de agitação .
There were five short-term studies that compared the effectiveness of Abilify and placebo over three weeks in a total of 1,900 **patients** .	Realizaram- se cinco estudos de curta duração que compararam a eficácia do Abilify com a de um placebo durante 3 semanas num total de 1 900 **doentes** .
Abilify is used to treat adults with the following mental illnesses : schizophrenia , a mental illness with a number of symptoms , including disorganised thinking and speech , hallucinations (hearing or seeing things that are not there) , suspiciousness and delusions (mistaken beliefs) ; bipolar I disorder , a mental illness in which **patients** have manic episodes (periods of abnormally high mood) , alternating with periods of normal mood .	O ABILIFY é utilizado no tratamento de doentes adultos com as seguintes doenças mentais : • esquizofrenia , uma doença mental com numerosos sintomas , incluindo pensamento e discurso incoerentes , alucinações (ouvir ou ver coisas que não existem) , desconfiança e ilusões (juízos errados) ; • perturbação bipolar I , uma doença mental em que os **doentes** sofrem episódios maníacos (períodos de humor muito elevado) , alternados de períodos de humor normal .

Os exemplos expostos acima foram coletados aleatoriamente com o objetivo de demonstrar como a palavra *patient* e seu correspondente tradutório *doente* se comportam em contextos de uso efetivo. Em português a palavra *patient* pode ser traduzida por *paciente*, assim como demonstra o dicionário online Michaelis:

Patient
n paciente. adj 1 paciente, perseverante. 2 resignado, conformado. 3 suscetível, apto a comportar. 4 passivo: que é objeto de uma ação.

Observa-se, todavia, que o termo em inglês que não é traduzido por *doente*. O exame de alguns contextos permite constatar que o item

lexical *doente* é efetivamente empregado na tradução de *paciente*, tal como nos exemplos discursivos que seguem:

1. "...uma doença mental em que os **doentes** sofrem episódios maníacos".
2. "Alguns **doentes** podem beneficiar de uma dose mais elevada".
3. "... durante 3 semanas num total de 1 900 **doentes."**

Nos casos acima citados, a palavra *doente* apresenta o mesmo significado de *paciente.* Embora os corpora utilizados para processamento e análise sejam essencialmente constituídos por textos do português europeu, acreditamos que em casos como o presente, deva haver concordância entre as duas modalidades (i.e. PE e PB).

Por meio da busca na ferramenta *online* Linguee, podemos perceber o uso do termo *patient*[28]*s* nas duas variantes linguísticas mais salientes na atualidade, isto é, o português europeu e o brasileiro. Observamos os seguintes exemplos:

Exemplo:

EN: [...] suggested system will be overly burdensome for competent authorities without leading to significant improvements in the quality of the information provided to patients. europa.eu[29]

PT: [...] seja excessivamente pesado para as autoridades competentes, sem resultar em melhorias significativas em termos de qualidade da informação fornecida aos doentes.

Exemplo:

EN: The impact has been so effective that currently in our institution the creatinine clearance is preventively evaluated prior to the administration of gadolinium to patients with moderate or severe renal insufficiency. rb.org.br[30]

[28] http://www.linguee.com.br/portugues-ingles/search?source=ingles&query=Patients

[29] Europa.eu is the official website of the European Union. It is a good starting point if you are looking for information and services provided by the EU but you don't know your way around our sites. http://europa.eu/abouteuropa/index_en.htm

[30] Radiologia Brasileira. http://www.rb.org.br

PT: O impacto foi tão efetivo que hoje já estamos calculando o clearance de creatinina antes de administrar o gadolínio em pacientes com insuficiência renal moderada ou grave.

Ao examinarmos os dois exemplos apresentados, torna-se evidente o correspondente **doente**, recuperado por meio do processamento automático relativamente ao termo *Patient*. Assim, a ferramenta parece demonstrar incialmente cumprir seu objetivo ao tratar da busca por correspondentes com nível de aceitabilidade confiável por meio do processamento automático.

2	MG	59493	0,589950144	1326	271633,0313

Por sua vez, a sigla *mg*, presente no quadro acima, que significa miligramas, apresenta chavicidade positiva no contexto medical. Ela aparece 59.493 vezes no corpus de estudo. Sendo assim, destaca-se como um termo importante podendo ser considerado candidato a termo no âmbito do campo disciplinar da área medical.

A busca por correspondentes tradutórios utilizando a ferramenta GIZA ++ (alinhamento de palavra por palavra) evidencia de forma pontual que *mg* (*milligram*) é traduzido por *mg*. (*miligramas*), conforme mostra abaixo:

CANDIDATO	CORRESPONDENTE	PROBABILIDADE
NULL	mg	0.0104685
Mg	mg	0.969518
mg/	mg/	1

Seguem abaixo exemplos de uso da sigla *mg* em inglês e seus correspondentes tradutórios em português.

TEXTO 1	TEXTO 2
It is available as 5 **mg** , 10 **mg** , 15 **mg** and 30 **mg** tablets , as 10 **mg** , 15 **mg** and 30 **mg** orodispersible tablets (tablets that dissolve in the mouth) , as an oral solution (1 **mg**/ ml) and as a solution for injection (7.5 **mg**/ ml) .	Está disponível na forma de comprimidos de 5 **mg** , 10 **mg** , 15 **mg** e 30 **mg** , de comprimidos orodispersíveis (que se dissolvem na boca) de 10 **mg** , 15 **mg** e 30 **mg** , sob a forma de uma solução oral (1 **mg**/ ml) e de uma solução injectável (7, 5 **mg**/ ml) .

For schizophrenia , the recommended starting dose is 10 or 15 **mg** by mouth per day .	Para a esquizofrenia , a dose inicial recomendada para o Abilify por via oral é de 10 ou 15 **mg** por dia .
The maintenance dose is 15 **mg** once a day , but higher doses may benefit some patients .	A dose de manutenção é de 15 **mg** uma vez por dia , mas alguns doentes podem beneficiar de doses superiores .
For bipolar disorder , the recommended starting dose is 15 **mg** by mouth once a day , either on its own or in combination with other medicines .	Para a perturbação bipolar , a dose inicial recomendada é de 15 **mg** por via oral uma vez por dia , tomada isoladamente ou em associação com outros medicamentos .
The solution for injection is only for short-term use and should be replaced by tablets , orodispersible tablets or oral solution as soon as possible : the usual dose is 9.75 **mg** as a single injection into the upper arm or buttock muscle , but effective doses range between 5.25 and 15 **mg** . A second injection can be given from two hours after the first if necessary , but no more than three injections should be given in any 24-hour period . 7 Westferry Circus , Canary Wharf , London E14 4HB , UK Tel .	A solução injectável destina- se apenas a uma utilização a curto prazo e deve ser substituída por 7 Westferry Circus , Canary Wharf , London E14 4HB , UK Tel .
The daily dose of Abilify should not exceed 30 **mg** , but this dose should be used with caution in patients who have severe problems with their liver .	A dose diária de Abilify não deve exceder 30 **mg** . Esta dose deve ser utilizada com precaução em doentes com problemas hepáticos graves .
In both studies of the solution for injection , patients receiving Abilify at doses of 5.25 , 9.75 or 15 **mg** had a significantly greater reduction in symptoms of	Em ambos os estudos efectuados com a solução injectável , os doentes a quem o Abilify foi administrado em doses de 5 **mg** , 25 **mg** , 9, 75 **mg** e 15 **mg** 2/ 3

agitation than those receiving placebo .	apresentaram uma redução significativamente superior nos sintomas de agitação , comparativamente com os doentes a quem foi administrado o placebo .
Injections of Abilify at doses of 10 or 15 **mg** were also more effective than placebo in reducing the symptoms of agitation , and were of similar effectiveness to lorazepam .	As injecções de Abilify em doses de 10 ou 15 **mg** demonstraram ser igualmente mais eficazes do que o placebo na redução dos sintomas de agitação , e demonstraram uma eficácia similar ao lorazepam .
5 **mg**	5 **mg**
5 **mg**	5 **mg**

Os exemplos acima evidenciam que *milligram* é frequentemente representado por sua abreviação *mg* (*miligrama*). Segundo a *Wikipédia*[31], "O miligrama (abreviado: *mg*) é uma unidade de massa do Sistema Internacional de Unidades". Um termo normalmente utilizado para identificar a medida (massa) de um comprimido na área da medicina. Trata-se de uma palavra representativa da área, pois no momento em que um médico prescreve um medicamento, caberá em geral também explicitar a dosagem indicada em termos de miligram. Tal prática parece se constituir como procedimento internacional. A proximidade tanto da palavra, quanto da sigla decorrente, acaba por gerar igualdade entre os termos e, por conseguinte, reduzindo os riscos de "erro" em sua tradução. Todavia, tal igualdade em um campo como o da medicina, precisa ser atestada e confirmada, sobretudo e inclusive através de sua verificação em contexto.

3	DOSE	49552	0,491372287	1640	222425,1875

O candidato a termo *dose* apresenta índice de chavicidade positivo e é representativo se comparado ao corpus de estudo e ao corpus de referência. Sendo assim, buscaram-se no GIZA ++ (alinhamento ao nível de palavra) os prováveis correspondentes de *dose*

[31] Disponível em maio de 2012 em: <http://pt.wikipedia.org/wiki/Miligrama>

e encontrou-se *dose* como correspondente tradutório em português. Observe-se no quadro que segue, abaixo:

CANDIDATO	CORRESPONDENTE	PROBABILIDADE
NULL	dose	0.000190862
Dosage	dose	0.0296769
fixed-dose	dose	0.0125762
Dose	dose	0.547485
dose-dependent	dose-	1

Em seguida apresentam-se linhas de concordâncias acompanhadas do contexto para o item, lexical *dose* em ambas as línguas, inglês e seus correspondentes tradutórios em português.

TEXTO 1	TEXTO 2
For schizophrenia , the recommended starting **dose** is 10 or 15 mg by mouth per day .	Para a esquizofrenia , a **dose** inicial recomendada para o Abilify por via oral é de 10 ou 15 mg por dia .
The maintenance **dose** is 15 mg once a day , but higher doses may benefit some patients .	A **dose** de manutenção é de 15 mg uma vez por dia , mas alguns doentes podem beneficiar de doses superiores .
For bipolar disorder , the recommended starting **dose** is 15 mg by mouth once a day , either on its own or in combination with other medicines .	Para a perturbação bipolar , a **dose** inicial recomendada é de 15 mg por via oral uma vez por dia , tomada isoladamente ou em associação com outros medicamentos .
Some patients may benefit from a higher **dose** .	Alguns doentes podem beneficiar de uma **dose** mais elevada .
To prevent manic episodes , the same **dose** should be continued .	Para prevenir episódios maníacos , a mesma **dose** deve ser continuada .
The solution for injection is only for short-term use and should be replaced by tablets , orodispersible tablets or oral solution as soon as possible : the	A solução injectável destina- se apenas a uma utilização a curto prazo e deve ser substituída por 7 Westferry Circus , Canary Wharf , London E14 4HB , UK Tel .

usual **dose** is 9.75 mg as a single injection into the upper arm or buttock muscle , but effective doses range between 5.25 and 15 mg . A second injection can be given from two hours after the first if necessary , but no more than three injections should be given in any 24-hour period . 7 Westferry Circus , Canary Wharf , London E14 4HB , UK Tel .	
The daily **dose** of Abilify should not exceed 30 mg , but this dose should be used with caution in patients who have severe problems with their liver .	A **dose** diária de Abilify não deve exceder 30 mg . Esta dose deve ser utilizada com precaução em doentes com problemas hepáticos graves .
The **dose** of Abilify should be adjusted in patients who are taking other medicines that are broken down in the same way as Abilify .	A **dose** deve ser ajustada para os doentes que estejam a tomar outros medicamentos metabolizados da mesma forma que o Abilify .
unit **dose** perforated blister (alu/ alu)	Blister de **dose** unitária perfurado (alu/ alu)
unit **dose** perforated blister (alu/ alu)	Blister de **dose** unitária perfurado (alu/ alu)

De acordo com os dados obtidos acima, o item lexical *dose* é um termo da área médica que significa *quantidade*. Observe os exemplos abaixo:

1. A **dose** *(quantidade)* diária de Abilify não deve exceder 30 mg .
2. A **dose** *(quantidade)* de manutenção é de 15 mg uma vez por dia...
3. Para a esquizofrenia , a **dose** *(quantidade)* inicial recomendada para o Abilify por via oral é de 10 ou 15 mg por dia .

Percebe-se que é perfeitamente possível substituir *dose* por *quantidade* sem perda do valor semântico do termo. Todavia, a palavra *dose* parece ser frequentemente utilizada na medicina como termo da área.

Em uma busca pela definição no dicionário monolíngue em língua inglesa *Longman Dictionary of Contemporary English Online*,[32] temos:

dose[33] [countable]
1 the amount of a medicine or a drug that you should take
dose of
Never exceed the recommended dose of painkillers.
high/low dose
Start with a low dose and increase it.

Podemos perceber na própria definição o uso da palavra *amount* se referindo à **quantidade** de. Nesse caso, referem-se a remédios utilizados para tratamento. Ao nos depararmos com o uso da palavra *exceed* no conjunto do contexto do termo *dose*, torna-se perceptível o claro uso do termo *dose* como quantidade.

Realizando uma busca em um dicionário monolíngue em língua inglesa, o Dicionário Priberam da Língua Portuguesa[34], e buscando pelo correspondente **dose**[35] em língua portuguesa temos a seguinte definição:

dose |ó|
(grego dósis, -eos, ato de dar)
s. f.
1. Quantidade determinada de cada ingrediente de uma preparação.
2. Porção de medicamento que se deve tomar de cada vez.
3. Quantidade, porção; ração.
dose de cavalo: a que é bem servida, em grande quantidade.
meia dose: (em restaurantes ou casas de pasto) quantidade de comida que se reputa suficiente para uma pessoa.
[Informal] ser dose: ser excessivo.
Confrontar: doce.

dosar - Conjugar
(dose + -ar)
v. tr.
1. Proceder à dosagem de.
2. Combinar nas devidas proporções.

[32] http://www.ldoceonline.com/
[33] http://www.ldoceonline.com/dictionary/dose_1
[34] http://www.priberam.pt
[35] http://www.priberam.pt/dlpo/default.aspx?pal=dose

Sinônimo Geral: DOSEAR, DOSIFICAR

No uso do termo **dose** em língua portuguesa percebe-se também a referência como **quantidade** e o aparecimento de um novo elemento: porção, que também está relacionado com o conceito de quantidade.

4	MEDICINAL	30001	0,297498792	263	140239,9375

O item lexical *medicinal* aparece 30.001 vezes no corpus de estudo e 263 vezes no corpus geral (corpus de referência), caracterizando uma palavra representativa e, logo, candidata a termo. Ao buscar o termo por meio da ferramenta *GIZA ++*, observa-se que o correspondente tradutório desponta como *medicamento,* tal como evidenciado abaixo:

CANDIDATO	CORRESPONDENTE	PROBABILIDADE
medicinal	medicamentos	0.533483

TEXTO 1	TEXTO 2
It explains how the Committee for **Medicinal** Products for Human Use (CHMP) assessed the studies performed , to reach their recommendations on how to use the medicine .	O seu objetivo é explicar o modo como o Comité dos **Medicamentos** para Uso Humano (CHMP) avaliou os estudos realizado , a fim de emitir recomendações sobre as condições de utilização do medicamento .
The Committee for **Medicinal** Products for Human Use (CHMP) decided that Abilify s benefits are greater than its risks for the treatment of schizophrenia and of moderate to severe manic episodes in bipolar I disorder , and for the prevention of a new manic episode in patients who experienced predominantly manic episodes and whose manic episodes responded to aripiprazole	O Comité dos **Medicamentos** para Uso Humano (CHMP) concluiu que os benefícios do Abilify são superiores aos seus riscos no tratamento da esquizofrenia e dos episódios maníacos moderados a graves na perturbação bipolar I e na prevenção de novos episódios maníacos nos doentes que experimentaram predominantemente episódios

treatment .	maníacos e em que os episódios maníacos responderam ao tratamento com aripiprazol .
NAME OF THE **MEDICINAL** PRODUCT	NOME DO **MEDICAMENTO**
Neuroleptic Malignant Syndrome (NMS) : NMS is a potentially fatal symptom complex associated with antipsychotic **medicinal** products .	Síndrome neuroléptico maligno (SNM) : o SNM é um conjunto de sintomas potencialmente fatal associado aos **medicamentos** antipsicóticos .
If a patient develops signs and symptoms indicative of NMS , or presents with unexplained high fever without additional clinical manifestations of NMS , all antipsychotic **medicinal** products , including ABILIFY , must be discontinued .	Se um doente desenvolver sinais e sintomas indicativos de SNM , ou apresentar febre elevada inexplicável sem manifestações clínicas adicionais de SNM , todos os **medicamentos** antipsicóticos deverão ser interrompidos , incluindo ABILIFY .
Lactose : patients with rare hereditary problems of galactose intolerance , the lapp lactase deficiency or glucose-galactose malabsorption should not take this **medicinal** product .	Lactose : os doentes com problemas hereditários raros de intolerância à galactose , deficiência lactase de Lapp , ou malabsorção glucose- galactose , não devem tomar este **medicamento** .
4 4.5 Interaction with other **medicinal** products and other forms of interaction	4. 5 Interacções **medicamentosas** e outras formas de interacção
Given the primary CNS effects of aripiprazole , caution should be used when aripiprazole is taken in combination with alcohol or other CNS **medicinal** products with overlapping undesirable effects such as sedation (see section 4.8) .	Atendendo aos efeitos primários do aripiprazol no SNC , deverá ter- se precaução quando o aripiprazol for administrado em associação com álcool ou outros **medicamentos** com acção no SNC e que tenham efeitos indesejáveis sobreponíveis , tais como a sedação (ver secção 4. 8) .
Potential for other **medicinal**	Potencial de outros

products to affect ABILIFY :	**medicamentos** para afectar ABILIFY :
Potential for ABILIFY to affect other **medicinal** products :	Potencial de ABILIFY para afectar outros **medicamentos** :

Ao analisar o candidato a termo *medicinal* observou-se que o mesmo é traduzido por *medicamento* e aparece prioritariamente acompanhado da palavra *products*. Observem-se os exemplos abaixo:

1. ... to affect other **medicinal products**.
2. Interaction with other **medicinal products**…
3. The Committee for **Medicinal Products** for Human Use…

Constatou-se, porém, que *medicamento* foi traduzido por *medicine*. Veja-se por exemplo: *... to reach their recommendations on how to use the **medicine***. Tradução: *... a fim de emitir recomendações sobre as condições de utilização do **medicamento***. Observou-se também que o dicionário online Michaelis[36] apresenta como correspondente tradutório de *medicinal* a palavra *medicinal* e *curativo*, como segue abaixo, reproduzido:

me.dic.i.nal
adj 1 medicinal. 2 curativo.

No dicionário online WorldReference[37] encontrou-se os seguintes correspondentes tradutórios:

medicinal adj (*healing*) medicinal, curativo adj
medicinal adj (*tasting like medicine*) que sabe a remédio medicinal, amargo adj

Diante dos resultados percebe-se que a palavra *medicinal* está relacionada a *medicamentos* quando acompanhada da palavra *products*, conforme evidências do corpus.

Ao realizarmos a busca na ferramenta online Linguee, pelo termo *medicinal*[38], encontramos usos do termo no contexto do português

[36]Disponível em maio de 2012 em:<http://michaelis.uol.com.br/moderno/ingles/index.php?lingua=ingles-portugues&palavra=medicinal>

[37] Disponível em maio de 2012 em: <http://www.wordreference.com/enpt/medicinal>

brasileiro, sem o acompanhamento da palavra *products* em que a tradução fosse realizada como medicina**is**.

Exemplo:

EN: For many years MISEREOR has promoted the cultivation of medicinal plants and vegetables in Belo Horizonte. misereor.org
PT: Desde muitos anos, MISEREOR fomenta o cultivo de plantas medicinais e hortaliças em Belo Horizonte. misereor.org

EN: Besides a description of the economic and cultural uses of many of these trees, guides also explain their medicinal value. ghanaembrasilia.org.br
PT: Além de uma descrição dos usos econômicos e culturais de muitas dessas árvores, os guias também explicam seu valor medicinal. ghanaembrasilia.org.br

Como pode ser observado com base nos exemplos e na recuperação do correspondente por meio da ferramenta *GIZA++*, *medicinal* é um termo que sofre mutações por meio de seus colocados.

Ao buscar pelo termo medicinal em um dicionário monolíngue em língua portuguesa encontramos uma definição que desconsidera os eventuais elementos que poderiam acompanhá-lo, tal como pode ser observado na fórmula: "Da medicina ou a ela relativo".

Observe-se, a seguir, a definição mais ampla, extraída do Dicionário Priberam da Língua Portuguesa em sua versão online[39]:

medicinal | adj. 2 g.
medicinal
(latim medicinalis, -e)
adj. 2 g.
1. Da medicina ou a ela relativo. = MÉDICO
2. Que serve de remédio.

Sendo assim, mesmo com a busca automática e recuperação de possíveis correspondentes tradutórios para esse termo, ao tradutor será sempre indicado consulta complementar nos corpora processados ou até

[38] http://www.linguee.com.br/portugues-ingles/search?source=ingles&query=medicinal
[39] http://www.priberam.pt/dlpo/default.aspx?pal=medicinal

mesmo em ferramentas de pesquisa em base de dados disponibilizadas *online* com a finalidade de encontrar correspondentes confiáveis.

É importante ter consciência de que a língua é um objeto abstrato, mas extremamente dinâmico, particularmente no que concerne a algumas de suas manifestações ligadas a compos específicos. As tentativas de congelar estados de língua, tal como o fez Saussure, pode responder perfeitamente às exigências para a instauração da Linguística, ou até mesmo para a realização de estudos científicos, cuja delimitação temporal se fizer necessária. Todavia, em geral, a tradução se confitura como uma atividade que considera a transferência de códigos baseada em textos. Por sua vez, os textos refletem atividades discursivas em constante mutação. Pode-se perfeitamente deduzir que muito embora a aplicação de técnicas digitais, realizadas com o auxílio de ferramentas de alta gama científica, possa oferecer resultados confiáveis, será importante considerar outras possibilidades instrutivas.

5	TREATMENT	39090	0,387627989	12124	132903,2969

Treatment é um termo frequente no corpus de estudo, ou seja, foi localizado 39.090 vezes, comparado ao corpus de referência que, a seu turno, apresentou 12.124 ocorrências. No corpus de geral (de referência) *treatment* pode ser considerado como *frequente*. Tal constatação nos indica que provavelmente *treatment* ocorre em textos de vários gêneros textuais. No corpus da área medicinal ele é considerado um candidato a termo, pois apresenta chavicidade positiva. Ao buscar a palavra *treatment* na ferramenta GIZA ++ (alinhamento paralelo ao nível da palavra) obteve-se como equivalente tradutório de *treatment* o item lexical *tratamento*. Observe-se abaixo:

CANDIDATO	CORRESPONDENTE	PROBABILIDADE
treatment	tratamento	0.868716

Os fenômenos encontrados por meio do alinhamento ao nível de palavras ficam elucidados nas linhas de concordância apresentadas na tabela abaixo:

TEXTO 1	TEXTO 2
If you need more information	Se necessitar de informação

about your medical condition or your **treatment** , read the Package Leaflet (also part of the EPAR) or contact your doctor or pharmacist .	adicional sobre a sua doença ou o **tratamento** , leia o Folheto Informativo (também parte do EPAR) ou contacte o seu médico ou farmacêutico .
For the **treatment** of schizophrenia , there were three main short-term studies of Abilify tablets lasting four to six weeks , which involved 1,203 patients and compared Abilify with placebo (a dummy treatment) .	No **tratamento** da esquizofrenia , foram efectuados três estudos principais de curta duração (quatro a seis 6 semanas) com Abilify na forma de comprimidos , que incluíram 1203 doentes e que compararam a eficácia do Abilify com a de um placebo (tratamento simulado) .
For the **treatment** of bipolar disorder , there were eight main studies looking at Abilify taken by mouth .	Para o **tratamento** da perturbação bipolar , foi igualmente avaliada a eficácia do Abilify tomado por via oral em oito estudos principais .
The eighth study looked at the effect of adding Abilify or placebo to existing **treatment** with lithium or valproate (another antipsychotic medicine) in 384 patients .	O oitavo estudo verificou a eficácia da adição do Abilify ou do placebo ao **tratamento** existente com lítio ou valproato (outro medicamento antipsicótico) em 384 doentes .
All of these studies looked at the change in symptoms using a standard scale for bipolar disorder or at the number of patients who responded to **treatment** .	Todos estes estudos analisaram a alteração dos sintomas utilizando uma escala padrão para a perturbação bipolar ou o número de doentes que responderam ao **tratamento** .
In the long-term studies , Abilify was more effective than placebo , and as effective as haloperidol , after up to a year of **treatment** .	Nos estudos de longa duração , o Abilify foi mais eficaz do que o placebo e tão eficaz quanto o haloperidol após um ano de **tratamento** .
Abilify was also more effective than placebo at preventing manic episodes returning in previously treated patients for up to 74 weeks , and when it was used as an add-	O Abilify demonstrou ainda ser mais eficaz do que o placebo na prevenção das recorrências de episódios maníacos em doentes tratados anteriormente durante um

on to existing **treatment** .	período de 74 semanas , e quando foi utilizado como adjuvante ao **tratamento** anterior .
The Committee for Medicinal Products for Human Use (CHMP) decided that Abilify s benefits are greater than its risks for the treatment of schizophrenia and of moderate to severe manic episodes in bipolar I disorder , and for the prevention of a new manic episode in patients who experienced predominantly manic episodes and whose manic episodes responded to aripiprazole **treatment** .	O Comité dos Medicamentos para Uso Humano (CHMP) concluiu que os benefícios do Abilify são superiores aos seus riscos no tratamento da esquizofrenia e dos episódios maníacos moderados a graves na perturbação bipolar I e na prevenção de novos episódios maníacos nos doentes que experimentaram predominantemente episódios maníacos e em que os episódios maníacos responderam ao **tratamento** com aripiprazol .
ABILIFY is indicated for the **treatment** of schizophrenia .	ABILIFY é indicado para o **tratamento** da esquizofrenia .
ABILIFY is indicated for the treatment of moderate to severe manic episodes in Bipolar I Disorder and for the prevention of a new manic episode in patients who experienced predominantly manic episodes and whose manic episodes responded to aripiprazole **treatment** (see section 5.1) .	ABILIFY é indicado para o tratamento do episódio maníaco moderado a grave na perturbação bipolar I e para a prevenção de novos episódios maníacos nos doentes que experimentaram predominantemente episódios maníacos e em que o episódio maníaco respondeu ao **tratamento** com aripiprazol (ver secção 5. 1) .

O texto acima mostra *treatment* e seu possível correspondente tradutório *tratamento* de forma contextualizada. Neste exemplo, a ferramenta GIZA ++ apresenta 0.868716 de probabilidade de a palavra *treatment* ser realmente correspondente tradutório de *tratamento*. Ao que indicam os exemplos contextualizados, a ferramenta *GIZA++* estaria correta.

6 PROPOSTA SEQUENCIAL PARA EXTRAÇÃO TERMINOLÓGICA PARALELA BILÍNGUE

Na proposta sequencial para extração terminológica bilíngue apresentada na investigação desenvolvida temos, resumidamente, as seguintes etapas:

1. Projeto do corpus;
2. Construção do corpus;
3. Processamento do corpus;
4. Extração dos candidatos a termos por meio das palavras-chave;
5. Extração dos correspondentes tradutórios para os candidatos a termos por meio do alinhamento ao nível de palavra;
6. Validação dos correspondentes tradutórios por meio de comparações como dicionários de uso geral ou de especialistas da área de conhecimento;

No que diz respeito ao projeto e construção do corpus, a proposta seguiu a abordagem já apresentada em outros trabalhos realizados e citados na introdução deste texto. No entanto, no que diz respeito ao processamento dos dados, a etapa de alinhamento ao nível de palavra foi apresentada como abordagem para automatizar o processo de recuperação de correspondentes tradutórios. Esta etapa até o presente momento mostrou-se eficaz diante dos objetivos estabelecidos para esta pesquisa.

Como pôde ser observado através dos dados obtidos, tanto relacionados à questão de probabilidade estatística, quanto às comparações com dicionários de uso geral como índice paralelo para validação da extração terminológica bilíngue e as linhas de concordâncias, a presente proposta sequencial pode constituir um meio interessante para a extração terminológica bilíngue em corpora paralelos – inglês/português – com vistas à tradução de textos das ciências médicas.

7 CONSIDERAÇÕES FINAIS

Diferentemente de alguns pesquisadores que, nesta sessão, falam de conclusões. Julgamos que à luz da ciência moderna só pode ser considerado como científico aquilo que pode ser contestado (cf. Demo, 2000). Logo, apesar de tratarmos de cálculos matemáticos, de probabilidades, é sempre importante modalizar a força dos resultados obtidos, tendo em vista que estamos a lidar com dados da linguagem, cuja natureza é essencialmente elástica, sobretudo em termos de sentido.

O presente trabalho visou oferecer uma proposta de ordem sequencial e criação de sistemas informáticos com vistas à extração terminológica bilíngue em corpora paralelos – inglês/português – voltada à tradução de textos das ciências médicas. Tal meta foi igualmente repetida na explanação do objetivo do trabalho.

Para levar a cabo a proposta, o estudo empregou ferramentas fornecidas pelo Processamento da Linguagem Natural, evocando principalmente as seguintes disciplinas: Linguística de Corpus, Corpora nos Estudos da Tradução, Terminologia e Extração Terminológica, com o intuito de oferecer um processo sistemático que contemplasse o processo de extração terminológica.

Os dados obtidos evidenciaram altos níveis de precisão, que levaram a supor que por meio da referida abordagem a recuperação de candidatos a termos e a busca por seus correspondentes tradutórios pode efetivamente ser otimizadas, revelando-se tão eficiente quanto a extração terminológica realizada analogicamente por especialistas. Em uma escala numérica de 0 a 1, a probabilidade de 0,822645962, 0.969518 e, em alguns casos, 1,0 (um), revelou a precisão dos correspondentes tradutórios.

As observações que seguem não pretendem prolongar o raio de ação estabelecido para este estudo. Visamos, através delas, tão somente refletir no escopo da questão, que ora nos conduziram a refletir sobre fatos indiretos ou cuja extensão convergirá para os estudos de corpora. Assim, salvo objetos concebidos como puro ornamento ou para mera satisfação ou contemplação artística, quase todos os apetrechos criados pelo ser humano para seu uso, buscam ampliar e superar suas limitações físico-biológicas e mentais, face ao próprio mundo artificial que o homem criou para si. Se para abrir latas de conservas é preciso uma lâmina especial, concebida para tal finalidade; se para percorrer uma distância de 500 km em apenas uma hora precisamos de um meio de transporte rápido. Porque razão não admitir, então, que para processar

milhões de palavras, centenas de textos, diante de metas e objetivos específicos, possamos empregar meios artificiais para fazê-lo?

Sabe-se que a recusa ou aversão à utilização de novas tecnologias é questão de adaptação e desejo de superar as supostas restrições que o uso da máquina exige. Quando existem, tais bloqueios decorrem, em geral, de ideias equivocada, que emergem dos choques inerentes às transformações. Trata-se, pois, no caso do processamento de textos, de questão de aceitação e especialização. Seus resultados, no caso do processamento de corpus, são atestações científicas empregáveis em prol da impossibilidade de poder fazê-lo *manualmente*, ou melhor, por meio do raciocínio.

O desaparecimento progressivo e parcial de algumas modalidades de consulta, ou mesmo de suporte à troca de informação, como o uso dos dicionários em versão papel, parece não oferecer saída possível, senão a adoção dos novos instrumentos. Quase sempre, os substitutos não são idênticos ou similares àqueles outrora empregados. Por exemplo, sabe-se que, atualmente, o uso de cartas enviadas por correio tradicional para trocar informações vem se reduzido progressiva e consideravelmente. Poderia se dizer o mesmo a respeito dos telefones convencionais (ditos fixos) que, pouco a pouco, vêm se tornando ornamentos ou peças de curiosidade para as crianças. Seguem a mesma linha as máquinas de datilografia e, de modo mais saliente, o próprio uso de lápis e canetas, que vem cada vez mais se reduzindo e sendo substituídos pela digitação em tela, sem que, por vezes, atentemos para tais modificações sociais.

A ortografia parece estar condenada a ceder lugar para a digitação, como aconteceu com a datilografia. Há cada vez menos mapas à venda nas livrarias e, por conseguinte, intensificação da venda de aparelhos GPS (*Global Positioning System,* Sistema de Posicionamento Global). Cada vez menos cálculos manuais para ceder lugar à prestação de valores prontamente calculados por computador. Até mesmo as calculadoras de mão tiveram sua utilização reduzida diante dos cálculos automáticos.

Com o advento dos leitores de livros, como o *Kindle*, e o consequente aumento das obras vendidas em versão *on-line*, supõe-se que em poucos anos será, talvez, possível não somente comprar e acessar textos de obras literárias para a leitura em tela, como também poder solicitar ao sistema versões da mesma obra traduzidas para outros idiomas, que, nesta projeção futurística, talvez estejam disponíveis nos mesmo pacote adquiridos.

Talvez seja possível vislumbrar, ainda, a possibilidade de solicitar que determinado sistema forneça estatísticas sobre palavras, expressões, localização de unidades com menções conceituais a lugares, personagens, etc. através de requisições simples e de forma interativa, tornando o estudo aqui apresentado um mero exemplo da evolução das técnicas aplicadas sobre os materiais que compõem as línguas.

Em suma, trata-se de projeções hipotéticas, mas nas quais os conhecimentos enciclopédicos poderiam estar ligados não somente aos componentes lexicais, como também a seus referentes conceituais atrelados às bases do texto. Por tais suposições, acredita-se que, embora ainda julgados demasiadamente mecanicistas pelo analista leigo, os trabalhos em Linguística de Corpus justificam o empenho que lhes vêm sendo dedicado por pesquisadores que integram análises linguísticas aos estudos discursivos de forma geral – por extensão: científicos, literários – aos resultados obtidos por meio o processamento automático de textos.

A crescente demanda pela realização de traduções científicas, em um grande número de áreas diferentes entre si, acentua a importância do processamento desses materiais e aperfeiçoamento dos suportes para fazê-lo (cf. Byrne, 2009). Provavelmente, tal premência científica deva-se não somente ao advento da globalização, mas também às tendências atuais, no sentido de que as ciências, cada vez mais, convergem para uma encruzilhada comum, aproximando saberes e fazendo com que os diálogos entre os diversos campos do conhecimento humano experimentem não somente acelerações, mas sobretudo entendimentos e acordos. Tais encontros entre as ciências se asseveram como inexoráveis e, por extensão, conduzem ao, igualmente inexorável, surgimento de novos termos, de neologismos e de inevitáveis cristalizações (temporárias) terminológicas, cujos desenvolvimentos fogem, por vezes, ao controle do próprio corpo crítico que as utiliza, sendo definido pelos utilizadores da língua.

Ademais, os saldos resultantes das integrações de componentes linguísticos para a composição de vocabulários específicos não possuem como fonte única a língua de base que sustenta sua manifestação. Por vezes, com algumas exceções em relação a línguas consideradas *francas*, como o inglês, parte dos termos adotados para tratar de campos específicos provém de outros idiomas, tornando os processos de tradução ainda mais complexos, pois nem sempre há normatizações definidas. Neste sentido, o processamento textual com vistas à extração dos itens presentes em corpora específicos contempla, atualmente, a identificação de absolutamente todos os elementos constantes nos corpora escritos. Parece ser evidente que os processos de quantificação,

de localização em contexto, de alinhamento, realizados de forma automática não forneçam exames qualitativos, mas tão somente parâmetros de suporte aos exames mais minuciosos. Como já observado várias vezes nos parágrafos que desenharam a presente tese de doutoramento, esta função constitui uma etapa a ser realizada *a posteriori* por meio de empenho humano. Muito embora os processamentos automáticos estejam cada vez mais próximos de fornecer informações qualitativas, o caráter quantitativo parece ainda preponderar.

Como apontado por Tebeaux (1997), o conhecimento científico se desenvolve em estreita relação com fatos linguísticos, antropológicos, culturais, históricos e políticos. Por extensão, estão também sujeitos às políticas linguísticas (e de línguas) que os regulam. A Aldeia Global, vislumbrada por McLuhan nos anos 1960 emergiu e se consolidou de forma mais incisiva após o falecimento do filósofo em 1980, pois as transformações sociais, sobretudo daquelas geradas pela revolução tecnológica, já despontavam no horizonte mesmo antes do século de Descartes, tendo se intensificado depois da revolução tecnológica gerada pelo surgimento dos primeiros computadores, lançados ao final da Segunda Grande Guerra. McLuhan foi capaz de prever parte dos impactos da tecnologia sobre as sociedades e sobre os comportamentos humanos. A partir de 1950, a força das telecomunicações não deixaria mais dúvidas a respeito das mudanças de paradigma que ocorreriam em relação às ideias anteriores sobre tempo, espaço, culturas e diversidades que envolvem o sujeito (*ego*, *nunc*, *hic*). A via IntelSatel, que marcou o envio de notícias dos exilados políticos da Europa e da América do Norte para o Brasil não se fecharia mais nas décadas posteriores, conduzindo inclusive à perda de controle em relação à proteção das seguranças nacionais. Em 2013, nem mesmo presidentes, forças de segurança, potências comerciais, estariam isentas do poder dos mecanismos de comunicação. A Aldeia Global, de McLuhan está fundada.

O que não se previu, tampouco se projetou, foi que apesar da homogeneização de comportamentos e, inclusive, da homogeneização que recai sobre as expressões estéticas, cujas manifestações, supõem-se, outrora seriam bem mais plurais, a humanidade não caminhou em direção à adoção de um código planetário único. Pelo contrário, as línguas hegemônicas, até o presente europeias, experimentaram flutuações e rodízios, a longo prazo, como centro do poder, mas permaneceram majoritárias, sobretudo como veículos de divulgação nos principais ramos da ciência.

Sabe-se, através do conhecimento histórico sobre as línguas, que o francês cedeu espaço ao inglês já a partir do século XVIII. O domínio anglofônico, por sua vez, em alguns setores parece já não ser mais hegemônico. Por exemplo, no próprio território americano o inglês progressivamente vem cedendo lugar ao espanhol em termos de número de falantes. A partir de 2013, países como a China vêm sendo despontando como potências dos próximos séculos (XXI ?). Se tais projeções se consolidarem, será de se esperar que outras línguas passem a ser foco de atenção e, provavelmente, se tornem os novos veículos de divulgação dos saberes científicos em decorrência das condições políticas que envolvem as posições hegemônicas.

Em suma, mesmo povos sem território definido, como os curdos e progressivamente os palestinos, parecem que continuarão a manter suas línguas em uso, paralelamente às línguas consideradas de prestígio. As diversidades linguísticas simplesmente se colocam à mercê das flutuações políticas. Mesmo que autores como Montgomery (2000) atestem a existência e predomínio das ditas línguas *francas* ou *universais*, em especial nas áreas científicas a demanda por traduções de textos, cujos originais se encontram expressos nas línguas ditas de *prestígio*, continuarão a ser requisitados em versão traduzida, justamente para garantir que os progressos científicos possam fortalecer as línguas nacionais, além de interesses primordialmente econômicos envolvendo outros povos.

Longe de se pautar como *lugar comum*, por constituir fato consubstancial à configuração dos modelos sócio-políticos da atualidade, o aperfeiçoamento dos processos de tradução, até o presente, continuam sendo realizados através das capacidades e habilidades decorrentes dos esforços cognitivos humanos. Destaca-se, todavia, a crescente importância concedida ao desenvolvimento de suportes, cujo acesso e manipulação propiciam ações mais ágeis ao tradutor (cf. Askehave, 2000). Como já observado, a tradução totalmente automática parece ainda estar longe de se tornar realidade, salvo casos específicos como o projeto canadense TAUM-Meteo, desenvolvido sobre microuniversos referenciais restritos e fechados que garante tradução perfeita a 100% no campo de boletins das condições do tempo.

Logo, do mesmo modo que se temeu, sobretudo durante os anos 1960, que a introdução de novas tecnologias no ensino – na época a televisão e o gravador de rolo – poderia substituir a figura do professor, sabe-se hoje, que isto não aconteceu e dificilmente acontecerá. Similarmente, o tradutor humano parece que continuará a ocupar por muitas décadas ou, provavelmente, ainda por centenas de anos, a base

central das atividades de interpretação, de tradução, especialmente no escopo poético-literário no qual a essência dos sentidos do verbo foge a qualquer tipo de apreensão de cunho matemático ou assertivo. Dessa forma, a tradução literária, sobretudo o trabalho poético, pensado aqui à luz dos postulados filosóficos de Barthes (2004) e Foucault (2002), autores que postulam o *apagamento do autor* ou a invisibilidade apontada por Venuti (1995), ou ainda metaforicamente como as obras de Galdi, jamais encontrarão fórmulas acabadas, ou seja, a tradução de texto literário implica competências que uma máquina dificilmente poderá realizar nos tempos atuais, ou seja, na data desta redação (2013).

Em revanche, não se pode negar que as máquinas se tornam – de forma impressionante –, cada vez mais, capazes de armazenar e preservar imensos volumes de dados, função outrora inimaginável mesmo por especialistas. Quantidades de informações nunca antes supostas estocadas em placas do tamanho de uma ervilha. Todavia, novamente tal certeza nos remete à quantidade e não ao exame científico-qualitativo, detalhado e respaldado com base na lógica humana, na razão, ou mais pontualmente: na criatividade.

Na comparação metafórica com um *cérebro* acumulativo-passivo a serviço dos seres inteligentes, um *computador* será capaz de reconhecer, de sintetizar e, até mesmo de imitar a fala humana, por extensão, poderá recitar, expor, dissecar, estratificar e, assim, se constituir como um aparato para a imitação e ampliação de habilidades e performances humanas, mas jamais analisar o verbo enquanto essência discursiva pautada sobre potências ideológicas de cunho político ou cultural. A ilha tecnológica de Bioy Casares, trata-a em sua obra intitulada A invenção de Morel, na qual o autor retoma as discussões sobre a mimeses, prega a morte do modelo e a vida autônoma da cópia. A tecnologia atual não se afasta das visões de Platão, quando o mesmo afirma que a vaca chorava no campo pelo bezerro de mármore que estava sendo transportado em uma carroça. a aproximação das representações digitais com as realidades já atingem graus impressionantes nas telas do cinema. Do mesmo modo, a Linguística de Corpus não se restringe mais ao tratamento de termos isolados, mas integrados ao discurso, conduzindo a impressão de que as possibilidades de processamento se aproximam da exaustividade, o que levaria à possibilidade de tradução de textos inclusive poético com base em estoques de realizações estocadas.

Pode-se aventar que a valorização do processamento do texto científico, bem como das técnicas para fazê-lo, só poderá gerar novos aportes científicos para a valorização do trabalho dos tradutores de

modo geral, sejam eles de orientação literária, estejam eles voltados ao trabalho com textos ditos de especialidade, como aqueles tratados nesta tese de doutoramento.

Como pôde ser observado na seção em que são apresentados os objetivos deste trabalho, foi possível testar a proposta de ordem sequencial apresentada para a realização de extração terminológica bilíngue inglês-português por meio do uso e do desenvolvimento de um conjunto de ferramentas oferecidas principalmente pela disciplina de linguística de corpus, em conjunto com o uso de corpora nos estudos da tradução e sua adequação ao contexto e execução do processo de extração terminológica.

Como ponto forte apresentado nesta pesquisa, o uso de recursos informáticos empregado até o presente momento de desenvolvimento do trabalho, tem sido central para a condução do processo de automatização de extração terminológica. A execução da proposta de ordem sequencial é concebida e analisada após cada processo computacional ser realizado. Embora todo o esforço empregado para o aperfeiçoamento tanto da proposta de ordem sequencial, quanto dos algoritmos computacionais utilizados tem-se em mente que é necessário considerar uma taxa mínima de erros (Och, 2003), mesmo tendo alcançado um progresso significativo ao longo do estabelecimento do uso de corpora em linguística e tradução, principalmente focada no processamento da linguagem natural.

Mesmo com o uso, integração e desenvolvimento das ferramentas utilizadas com a finalidade de aperfeiçoamento do processo de extração terminológica bilíngue aqui, citamos o uso de ferramentas desenvolvidas para fins específicos, como por exemplo, o *WordSmith Tools*, e que são, geralmente, comerciais de forma que não temos acesso ao código fonte para implementação em um sistema automatizado que contribuíram para certificar que os cálculos empregados e os algoritmos desenvolvidos são eficientes e produzem resultados confiáveis se comparados àqueles resultantes do processamento em programas comerciais.

Dentre as ferramentas utilizadas, percebeu-se a importância dos recursos de *WordList* (lista de palavras) e *KeyWords* (palavras-chave), presentes na suíte de ferramentas do *WordSmith Tools* para a obtenção da lista de palavras e da lista de palavras-chave. Essas ferramentas serviram como base para certificação de que os dados gerados por aqueles algoritmos desenvolvidos para este projeto, o qual que passou a contar com ferramentas próprias de geração de listas de palavras e palavras-chave, foi possível gerar resultados tão precisos quantos

aqueles gerados por ferramentas comerciais. Outra ferramenta de grande impacto para o sucesso das etapas realizadas até aqui foi a *GIZA++*. Por proporcionar o alinhamento ao nível de palavra esta ferramenta serviu como base para a composição automática da lista de correspondentes tradutórios dos itens lexicais obtidos por meio do índice de chavicidade produzidos através do cálculo de palavras-chave. Entre outras ferramentas, destacam-se ainda aquelas que serviram para visualização dos dados obtidos.

No avanço da pesquisa, as próximas etapas buscaram desenvolver ainda um sistema de disponibilização do glossário bilíngue inglês-português e a possibilidade de exportação para os formatos de programas de apoio à tradução largamente utilizada por profissionais da tradução.

Seguindo recomendações apontadas pelo pesquisador Tony Beber, que participou da etapa de qualificação desta pesquisa, procurou-se centrar esforços na integração entre a ferramenta *WordSmith Tools* e *GIZA++* com a finalidade de demonstrar aos pesquisadores e usuários de ferramentas de extração terminológica seu uso e resultados que pode ser obtidos. Mesmo que a pesquisa tenha contemplado a criação de ferramentas próprias, destacamos que com uma integração projetada das ferramentas atualmente disponíveis aos usuários finais, podemos conseguir resultados satisfatórios.

Os resultados apresentados por meio do processamento dos corpora compilados puderam demonstrar que os índices de probabilidade dos correspondentes tradutórios foram elevados. Fato que prova a significância da proposta desenvolvida nesta pesquisa. Isso nos possibilitou verificar significativos avanços tecnológicos na área da tradução. Mesmo que ainda não existam máquinas de traduções satisfatórias, parece que estamos caminhando a passos largos no seu desenvolvimento.

Situados em um universo com uma alta demanda de traduções, principalmente, em áreas científicas, os instrumentos de auxílio à tradução têm como principal objetivo proporcionar as intérpretes e tradutores instrumentos e mecanismos que aperfeiçoem sua produtividade, além de promover o alto índice de qualidade durante.

Uma das singularidades na proposta apresentada foi o fato de preencher as lacunas por pesquisas até então realizadas no âmbito da utilização de ferramentas de corpora para a extração terminológica. Como visto em Portolan (2011), por exemplo, entre outros já citados no decorrer desta pesquisa, para a elaboração de um glossário bilíngue na área de pediatria, com proposta principal a autora cita como objetivo

geral "Contribuir para a elaboração de um glossário bilíngue português/inglês na área de Pediatria", no entanto, na busca por correspondentes a recuperação é feita de forma semiautomatizada por meio da comparação de listas de palavras-chave entre corpora de línguas distintas.

Além do percurso realizado por Portolan (Ibid.) em sua proposta, outros pesquisadores ainda citaram a dificuldade em conseguir realizar a busca por correspondentes tradutórios aqueles candidatos a termos previamente processados pelos algoritmos de geração de palavras-chave ao aplicar estatísticas de cálculo de qui-quadrado ou razão de verossimilhança. Um ano posterior ao trabalho de Portolan (Ibid.), Foo (2012) afirma que:

> Outro problema na criação de terminologia é encontrar relações entre candidatos a termo. A saída mais comum a partir de algoritmos ATE é uma simples lista de candidatos a termos ordenados. Em uma lista desse tipo, é difícil para o terminológo examinar termos relacionados que devem ser considerados ao realizar a análise de conceito. A lista também é uma representação inadequada quando é necessário movimentar-se entre os diferentes níveis de granularidade de informação para grandes conjuntos de dados. (FOO, 2012, p. 59, tradução nossa)[40]

Com a evolução computacional, podemos ir além da comparação de listas de palavras por promover a busca automatizada baseado em algoritmos específicos para tal tarefa. No caso desta pesquisa procurou-se demostrar, de forma exaustiva, a utilização da ferramenta GIZA++ como base para a busca dos correspondentes tradutórios.

Com a integração das diferentes ferramentas para a criação de uma de ordem sequencial e criação de sistemas informáticos para extração terminológica bilíngue em corpora paralelos – inglês/português – com vistas à tradução de textos das ciências médicas, chegamos a um

[40] Another issue in terminology creation is finding relationships between term candidates. The most common output from ATE algorithms is a plain list of ordered term candidates. In such a list, it is difficult for the terminologist to examine related terms which must be considered when performing concept analysis. A list is also an ill-suited representation when it is necessary to move between different levels of information granularity for large data sets.

ambiente semelhante aquele que Rettig, Simons e Thomson (1993) criaram e chamaram de *CELLAR* (*Computing Environment for Linguistic, Literary, and Anthropological Research*), ou seja, uma plataforma para pesquisas nas áreas da linguística, literatura e antropologia. Baseado na proposta dos autores, colocamos a proposta aqui apresenta como uma ferramenta que contém os requerimentos para o manejo dos dados linguísticos.

Para Simons (1998), são seis os requerimentos apontados pelo autor para se criar um ambiente de processamento linguístico:

1. Os dados são multilíngues, assim o ambiente de computação deve ser capaz de rastrear em quais dados pertence a cada linguagem, e, em seguida, exibir e processá-lo em conformidade.

2. Os dados em texto se desdobram sequencialmente, de modo que o ambiente de computação deve ser capaz de representar o texto na sequência adequada.

3. Os dados são estruturados hierarquicamente, de modo que o ambiente de computação deve ser capaz de construir estruturas hierárquicas de profundidade arbitrária.

4. Os dados são multidimensionais, para o ambiente de computação tem de ser capaz de unir diversos tipos de análise e interpretação de dados únicos.

5. Os dados são altamente integrados, de modo que o ambiente de computação deve ser capaz de armazenar e seguir os vínculos associativos entre as partes dos dados relacionados.

6. Ao fazer todos os itens acima para modelar a estrutura de informação dos dados corretamente, o ambiente de computação deve ser capaz de exibir dados convencionalmente formatados.

> (SIMONS, 1998, p. 23-24, tradução nossa)[41]

Com base nos seis requerimentos apresentados por Simons (Ibid., p. 23-24) a proposta comporta dados multilíngues sendo capaz de processar os mesmos e produzir resultados adequados, não somente no par inglês /português. Os dados são processados de forma hierárquica e altamente integrados e em formatos convencionais de exibição, podendo ser exportados para diversos outros formatos proprietários ou não, também podendo ser integrado em bancos de dados linguísticos, como, por exemplo, aquele apresentado por Fromm (2007) na criação do Votec, um sistema para a construção de vocabulários eletrônicos para aprendizes de tradução.

De acordo com Fromm (Ibid.), expomos aqui em suas palavras a possibilidade de integração:

> A opção da construção de um banco de dados em um padrão que pode ser facilmente alterado (de acordo com a documentação constituída para o mesmo) não foi aleatória: novas ferramentas podem ser acrescidas para automatizar, sempre que possível, partes do processo. Muitos pesquisadores já se dedicam a criar essas ferramentas e os exemplos mais bem acabados são aquelas desenhadas para a extração automática de corpus e, a partir desse, da extração automática de candidatos a termos. (FROMM, 2007, p. 161-162)

---

1. [41] The data are multilingual, so the computing environment must be able to keep track of what language each datum is in, and then display and process it accordingly.

The data in text unfold sequentially, so the computing environment must be able to represent the text in proper sequence.

2. The data are hierarchically structured, so the computing environment must be able to build hierarchical structures of arbitrary depth.
3. The data are multidimensional, so the computing environment must be able to attach many kinds of analysis and interpretation to a single
4. datum.
5. The data are highly integrated, so the computing environment must be able to store and follow associative links between related pieces of data.
6. While doing all of the above to model the information structure of the data correctly, the computing environment must be able to present conventionally formatted displays of the data.

Com relação ao rápido avanço e evolução da tecnologia, o autor menciona com propriedade a abertura para integração futura com sua ferramenta:

> Devido aos constantes avanços da técnica, deve-se considerar esta proposta como um ponto de partida para novas empreitadas na área que combinem Inteligência Artificial, Mineração de Dados e criação de Ontologias e Taxonomias. (FROMM, 2007, p. 162)

Desta forma, a proposta abordada neste trabalho fornece ferramentas necessárias para a exportação e/ou integração que contemplam diversos trabalhos já realizados por pesquisadores com a preocupação de produzir recursos terminológicos tanto para pesquisa, quanto para auxilio a tradutores profissionais. Cabe mencionar que no caso de uso da proposta apresentada por realizar a busca por termos e correspondentes de forma automatizada, tem o principal objetivo de fornecer recursos de forma rápida para a atividade tradutória.

Destaca-se, ainda, o uso de corpora de grande escala para a aplicação da proposta abordada ate aqui, pois os mesmos podem influenciar diretamente nos resultados para a busca por candidatos a termos e seus correspondentes tradutórios. Isso é valido para uma gama de pesquisa com a utilização de corpora. Para Hoard (1998), destacam-se alguns usos com resultados satisfatórios no uso de corpora de grande escala, são eles:

> (1) usando a indexação conceitual de um grande número de textos curtos (ou textos longos segmentados em "pedaços" mais adequados), para selecionar textos sobre temas específicos para algum propósito linguístico ou outros, (2) selecionar exemplos de conjuntos de alguns fenômenos linguísticos em coleções de textos muito grandes, e (3) encontrar equivalentes bilíngues de itens lexicais em (presumivelmente) contextos equivalentes. (HOARD, 1998, p. 226)[42]

[42] (1) using conceptual indexing of a large number of short texts (or long texts segmented into suitable “chunks”), to select texts on particular topics for some linguistic purpose or other, (2)

Em conformidade com Hoard (Ibid.), demos maior importância para o item de número 3 (três) apontado pelo autor. Da mesma forma que ele cita que com o uso de corpora de grande escala resultados podem ser melhores na busca por equivalentes bilíngues, as ferramentas empregadas nessa proposta se baseiam fortemente no tamanho dos corpora processados na busca por melhores resultados.

Berber Sardinha (2002), em seu artigo intitulado Tamanho do Corpus, define a escala de tamanho de corpora como reproduzido na Tabela 5, abaixo.

Tabela 5: Classificação relativa ao tamanho de corpus (Berber Sardinha, 2002, p. 119)

TAMANHO EM PALAVRAS	**CLASSIFICAÇÃO**
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Como pode ser observada na tabela, a proposta realizada nessa pesquisa fez uso de corpora de tamanho grande, como definido por Berber Sardinha (Ibid.) e tem corroborado com os requerimentos apontados por Hoard (1998) e por isso pudemos ter êxito nos resultados obtidos através do processamento dos corpora compilados. Traduzindo em números temos corpora processados para essa análise contendo um total de 12.459.878 (doze milhões, quatrocentos e cinquenta e nove mil, oitocentos e setenta e oito) palavras no texto em língua inglesa e 15.653.159 (quinze milhões, seiscentos e cinquenta e três mil, cento e cinquenta e nove) no texto em língua portuguesa.

Por fim, fazendo alusão à escala apresentada por Berber Sardinha (Ibid., p. 119) demonstramos ainda o posicionamento dos corpora utilizados para a testagem da proposta de ordem sequencial e criação de sistemas informáticos para extração terminológica bilíngue em corpora paralelos – inglês/português – com vistas à tradução de textos das ciências médicas, na Figura 21, a seguir.

---

culling example sets of some linguistic phenomena from very large collections of text, and (3) finding bilingual equivalents of lexical items in (presumably) equivalent contexts.

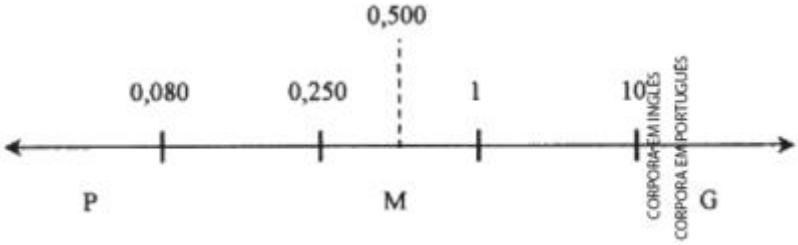

Figura 21: Escala de tamanho de corpora baseada em Berber Sardinha 2002, p. 119)

Baseado nos itens apontados anteriormente como os requerimentos, a integração e os resultados confiáveis relacionados ao tamanho dos corpora empregados, a pesquisa pode apresentar resultados do quais aprovam a proposta até então desenhada.

A proposta de extração terminológica apresentada aqui se aplica tanto em corpora de tamanhos pequenos, quanto em corpora de grande escala. Cita-se o fato de que quanto maior o tamanho dos corpora, em número de textos ou palavras, resultados mais significativos deverão ser alcançados. Além disso, concordamos com a colocação de Almeida et. al. (2006) que diz:

> [...] percebe-se que a associação entre terminologia e informática é viável e, sobretudo, necessária para as ações e pesquisa de soluções terminológicas assistidas por computador. Soma-se a isso a existência de ferramentas e/ou ambientes para o português, o que confere às pesquisas terminológicas em língua portuguesa um avanço evidente. (ALMEIDA et. al., p. 44)

Krieger (2009, p. 45) também, ao salientar o papel da terminologia na construção dos saberes científicos, cita a técnica e a tecnologia como designativas de componentes. Cada vez mais temos fenômenos da globalização acontecendo. Em diversas áreas, todas essas especificas, há a necessidade por tradução em línguas especializadas a ser realizada a velocidade da luz, uma vez que essas favorecem a economia, os negócios, intercâmbios e etc. cria-se também a necessidade pela extração terminológica na mesma velocidade. A autora

cita a importância dos bancos terminológicos, glossários e dicionários técnico-científicos por fazerem parte de campos do saber bastantes específicos:

> O valor esse tipo de obra de referência vincula-se ainda ao fato de que, mesmo indiretamente, todas cumprem um papel normalizador, vale dizer, constituem-se em instâncias que determinam a norma, o padrão linguístico e conceitual a ser corretamente empregado. (KRIEGER, 2006, p. 46)

Seguindo a linha de pensamento da autora, enquanto tivermos um intercambio de diversos itens entre as línguas como, por exemplo, comércio entre países de línguas diferentes, por meio da importação e exportação, e o mesmo depender da comunicação teremos que ter ferramentas preparadas para suprir essas demandas cada vez mais urgentes.

Já mencionada à importância da terminologia no âmbito dessa pesquisa, vela a pena destacar novamente sua importância e relacionar com a relevância da proposta desenvolvida aqui, uma ferramenta que permitirá aos especialistas o processamento rápido e eficaz suprindo necessidades do fenômeno da globalização iniciado no final do século XX e que é fundamental para a economia mundial no início e decorrer do século XXI. De acordo com o *Pointer Final Report* (1996), vemos o papel fundamental da terminologia:

> Uma vez que uma grande parte desta comunicação - especializada - baseia-se no vocabulário de um vasto número de áreas sujeitas a transmitir seu conteúdo, facilmente acessível, a terminologia irá desempenhar um papel cada vez mais importante na gestão da informação (multilíngue) no século XXI. (DIT, 1996, tradução nossa)[43]

[43] Since a great deal of this - specialist - communication relies on the vocabulary of a vast number of subject fields to convey its content, readily-accessible, up-to-date terminologywill play an increasingly important role in (multilingual) information management in the 21st century.

A proposta, mesmo que faz o uso de ferramentas já conhecidas por estudiosos da linguística, terminologia e tradução, teve o objetivo de apontar novas possibilidades que facilitaram e fornecerão subsídios para melhorias em propostas futuras.

Na pesquisa aqui desenvolvida pudemos ter acesso a criação de algoritmos para a geração de listas de palavras e listas de palavras-chave que é largamente usado na ferramenta *WordSmith Tools*, tornando essa geração independente de programas de terceiros dentro de uma proposta. A suíte de ferramentas disposta pela GIZA++ é distribuída com licença de código aberto, isso faz com que alterações necessárias para a integração em uma proposta em que tenha como objetivo um sistema autônomo, completamente independente de plataforma e pagamentos de taxas[44] (preço pago pela ferramenta *WordSmith Tools*).

Respondendo os questionamentos iniciais desta pesquisa, para a primeira pergunta: (a) o processo de busca por correspondentes tradutórios, posto em operação através do uso de corpora poderia ser alcançado através da utilização das tecnologias que se têm atualmente disponíveis, ou seria necessário acrescentar programas que possam gerar dados de naturezas diferentes daqueles comumente obtidos? Pelo decorrer dos testes e análises dos dados, com o uso de tecnologias já disponíveis foi possível alcançar resultados satisfatórios para a extração terminológica.

O uso de ferramentas como o *WordSmith Tools* e o conjunto de ferramentas como o GIZA++ na integração da proposta, mesmo usando de métodos já bastante mencionados em trabalhos, mas não de forma integrada ainda podem fornecer subsídios para extração terminológica. No entanto, trabalhos que procuram mostrar uma sequência de etapas para a extração terminológica automatizada ainda são escassos. Desta forma, esta pesquisa contribui para uma discussão na busca por uma proposta definida para tal tarefa.

Outro questionamento que podemos responder ao concluir essa tese de doutoramento está ligado na seguinte questão: (b) qual a razão para que estudos já realizados se baseiem nas mesmas etapas metodológicas, porém não debatam a questão de busca por correspondentes tradutórios de forma automatizada? Tal tarefa seria realizável? Caso positivo, qual o grau de precisão e de confiabilidade dos resultados relativamente às necessidades do tradutor? Respondendo esses questionamentos em segmentos, acredita-se que a razão pela qual

[44] Nome do produto: WordSmith Tools Single User [#206310] Preço: BRL 199.52 em https://secure.shareit.com/shareit/cart.html?PRODUCT[206310]=1&js=-1

os estudos já realizados sobre extração terminológica não contemplem a busca por correspondentes tradutórios de forma automatizada esteja ligada principalmente ao escopo das pesquisas. Os pesquisadores seguem uma tendência de realizar a busca por candidatos a termos por meio de programas de geração de palavras-chave como, por exemplo, *WordSmith Tools, AntConc*[45] etc. e em sequência realizam um relacionamento manual entre as lista de palavras-chave nos diversos idiomas sendo, geralmente, em inglês e português.

Para a segunda subquestão, a tarefa de busca por correspondentes tradutórios para os candidatos a termos mostrou-se totalmente viável por meio da utilização de ferramentas adequadas. Com o auxílio do conjunto de ferramentas fornecido pelo GIZA++, um grande passo foi dado por calcular automaticamente a relação entre os correspondentes tradutórios.

Além dos recursos principais do GIZA++, ele ainda fornece uma vasta gama de extensões que podem melhorar os resultados na busca por correspondentes tradutórios, dentre elas, podemos citar: alinhamento multissegmentado, alinhamento forçado, alinhamento retomado e ainda outro recurso em que é possível realizar alinhamento parcial (manual). Para uma visão detalhada desses recursos e sua utilização, conferir o artigo *A Semi-supervised Word Alignment Algorithm with Partial Manual Alignments*[46] de Qin Gao, Nguyen Bach e Stephan Vogel. Para um tutorial completo na utilização dessas ferramentas, conferir *MGIZA*[47].

Como todos esses benefícios apresentados por um único conjunto de ferramentas percebe-se que é possível integrar essa ferramenta nessa proposta, como também em propostas já apresentadas anteriormente por pesquisadores com a finalidade de promover a extração terminológica com o auxílio de dispositivos computacionais.

---

[45] http://www.antlab.sci.waseda.ac.jp/software.html

[46] http://www.cs.cmu.edu/~nbach/papers/acl-wmt10.pdf

[47] MGIZA++ is a multi-threaded word alignment tool based on GIZA++. It extends GIZA++ in multiple ways: Multi-threading: MGIZA++ can make use of multi-core platforms efficiently. Usually a quad-core machine can have a three-fold speedup over single-thread GIZA++. Memory optimization: By eliminating duplicated tables, MGIZA++ can save a lot of memory comparing to GIZA++. Resume training: MGIZA++ can resume training from any stage and continue training. For example you may be able to re-use previous available models and continue training directly from IBM Model 4 instead of all the way from Model 1. Integrated with Chaski: MGIZA++ can be integrated into Chaski and run on cluters, which will give you even larger speedup. Native Windows support: MGIZA++ can now be compiled in Visual Studio, providing native MS Windows support. The latest version is, however, not stable when compiled as 64bit. http://www.kyloo.net/software/doku.php/mgiza:overview

Finalizando esse questionamento, ao responder sua última subquestão, tem-se como positivo a integração da busca por correspondentes tradutórios de forma automatizada, em que pudemos verificar que o grau de precisão e de confiabilidade dos resultados obtidos por meio do processamento automático foi satisfatório quando comparado com dicionários e outros materiais de alta relevância.

Esses resultados podem ser visualizados na seção 5 DADOS: ANÁLISE. Em um índice entre 0 e 1, percebeu-se que grande parte dos resultados para a lista de candidatos a termos chegou a um patamar acima de 50%, ou seja, 0,5. Nota-se em alguns exemplos retomados:

CANDIDATO	CORRESPONDENTE	PROBABILIDADE
patients	doentes	0.881883
treatment	tratamento	0.868716
medicinal	medicamentos	0.533483

O último questionamento levantado para essa pesquisa, fortemente baseado no parecer apontado pelo professor Tony Berber Berber Sardinha[48] na etapa de qualificação, em que reformulamos a pergunta em razão dos apontamentos para: c) resultados satisfatórios podem ser alcançados por meio dos programas computacionais, no caso dessa pesquisa a ferramenta *GIZA++* em conjunto com o *WordSmith Tools*, fornecer auxílio ao tradutor em sua necessidade de terminologia ad hoc?

Já discutido ao responder os questionamentos anteriores, pode-se ser percebido que a integração dos programas *WordSmith Tools* e *GIZA++* supriram as necessidades pontuais da pesquisa de doutoramento com êxito. Foi possível nas seções 3 FUNDAMENTAÇÃO METODOLÓGICA, 3.1 APERFEIÇOAMENTOS AOS SUPORTES INFORMÁTICOS, 4 METODOLOGIA: PROJETO, CONSTRUÇÃO E PROCESSAMENTO DOS CORPORA, 4.2.6 Ferramenta plain2snt.out, 4.2.7 Ferramenta mkcls, 4.2.8 Conjunto de Ferramentas GIZA++ e 4.3.1 Ferramentas de Processamento do Corpus demonstrar sua integração e uso como exemplos que podem ser seguidos e aprimorados por pesquisadores e especialista da área da terminologia, com também pelos usuários finais, os tradutores.

[48] http://www2.lael.pucsp.br/~tony/tony/Home.html

Ao retomar os objetivos dessa tese de doutoramento, em que se visou propor uma ordem sequencial de aplicação de programas computacionais para tratar da questão da extração terminológica semiautomática por meio de ferramentas de análise estatística em corpora paralelos – inglês/português – destinados à geração de glossários bilíngues, com vistas à tradução de textos das ciências médicas, tem-se a percepção que o mesmo foi alcançado com êxito ao desenhar uma proposta que apresentou dados empíricos no processamento dos corpora e os resultados obtidos comprovaram sua eficiência. Cita-se o fato que a proposta pode ser alterada em qualquer de suas etapas com a finalidade de atender requisitos de futuras pesquisas baseados em seu escopo e objetivos. O desenho da proposta completa pode ser encontrado de forma resumida na seção 6 PROPOSTA SEQUENCIAL PARA EXTRAÇÃO TERMINOLÓGICA PARALELA BILÍNGUE.

Já com relação aos objetivos específicos, retomamos um a um coma finalidade de rever o potencial da pesquisa realizada. Com relação ao primeiro: desenvolver uma proposta para a extração de glossários bilíngues a partir de corpora paralelos, visando preencher supostos gaps (lacunas) passíveis de aperfeiçoamento, com vistas ao processamento e à extração terminológica para a geração de glossários bilíngues, foi conseguido realizar e demonstrado em seu desenho completo, descrevendo todas suas etapas. Esse primeiro objetivo específico está fortemente relacionado ao objetivo geral da pesquisa e por isso sua resposta é assertiva.

Em relação ao segundo objetivo específico: elaborar programas computacionais específicos para a proposta de ordem sequencial como, por exemplo, gerador de listas de palavras, listas de palavras chaves por meio das abordagens de qui-quadrado, correção do qui-quadrado e da razão de verossimilhança de modo a atingir resultados mais precisos, a pesquisa pode discutir e desenvolver na ferramenta *WordStats* os cálculos estatísticos tanto na ordem do qui-quadrado, correção do qui-quadrado e da razão de verossimilhança. No entanto, em conformidade com o terceiro questionamento de pesquisa, apresentamos no relatório os dados obtidos por meio do programa *WordSmith Tools*.

Por fim, o terceiro e ultimo objetivo específico: analisar qualitativamente parcela dos resultados obtidos em relação ao par de línguas trabalhado com a finalidade de validar os termos recuperados de forma automática pela máquina, foram realizados e demonstrados na seção 5 DADOS: ANÁLISE e puderam validar a proposta da pesquisa por apresentarem resultados satisfatórios em suas análises.

Como pontos fortes de contribuição desta pesquisa e que puderam ser discutidos pode-se citar:

a) a apresentação e a discussão de proposta de uma ordem sequencial completa de um sistema de extração terminológica por meio de corpora paralelos bilíngues devidamente alinhados;

b) identificação e proposta de solução para o problema ligado a busca de correspondentes tradutórios de forma automatizada ao empregar ferramentas de alinhamento ao nível de palavra;

c) apresentação de um arcabouço de ferramentas utilizadas em uma ordem específica para a finalidade de geração automática de glossários bilíngues que tem a possibilidade de ser aperfeiçoado com a inclusão, exclusão, ou remodelagem da ordem de procedimentos envolvidos e passível de ser utilizada para diferentes pares de línguas.

Tem-se em mente que com o desenvolvimento desta proposta sequencial e com os resultados obtidos por ela, lacunas que permeiam a extração terminológica de forma automatizada, principalmente centrada na busca por correspondentes tradutórios sem a necessidade de exame ou comparação humana, puderam ser preenchidos. No entanto, mesmo com o sucesso obtido através da proposta sequencial utilizada, destacamos que recorrer à peritos e profissionais da área científica em cujo escopo a extração terminológica foi realizada é indispensável para a certificação dos termos e candidatos a correspondentes tradutórios. Tal etapa não foi realizada por questão de tempo. Todavia, pretende-se prolongar a presente pesquisa neste sentido.

Cabe ainda salientar que a proposta de extração terminológica apresentada aqui se aplica tanto em corpora de tamanhos pequenos, quanto em corpora de grande escala. Cita-se o fato de que quanto maior o tamanho dos corpora, em número de textos ou palavras, resultados mais significativos deverão ser alcançados.

Para trabalhos futuros, ou mesmo a continuidade desta pesquisa em outro nível, destacam-se alguns pontos que poderiam ser mais bem discutidos e trabalhados. Entre eles, como principal função no auxílio aos tradutores e especialistas da terminologia tem-se a questão de termos compostos. Em um relato baseado em Foo (2012, p. 58) focado nos estudos em extração terminológica os autores citam trabalhos de Foo e Merkel (2010) que, por exemplo, no caso da língua sueca “há

mais termos simples do que compostos"[49], neste estudo os autores destacam que aproximadamente um terço (exatamente 27.95%) dos candidatos a termos coletados são termos compostos. Ainda, Foo (Ibid.) cita o trabalho de Lefever et. al. (2009) em que ao trabalhar com o par linguístico francês-alemão na realização da extração terminológica bilíngue destaca o fato que mesmo decompondo os termos compostos extraídos, abordagem se faz correta. No caso do par linguístico francês-alemão, o autor cita que um dos motivos pelos quais isso se faz verdadeiro é o fato que "na língua francesa palavras composta são relativamente raras".

Outro ponto que vale destacar ao finalizar essa pesquisa de doutoramento e se pensar na abertura de seus resultados para consulta direta ao público em geral são a criação de uma interface gráfica para acesso aos dados processados, ou mesmo seu processamento *Q.E.D.* Esse item apresentaria uma visualização dos candidatos a termos bilíngues recém-extraídos, também listagem dos possíveis correspondentes tradutórios para cada um dos candidatos a termo.

Junto com uma interface gráfica para apresentação dos termos e seus respectivos correspondentes tradutórios, remete-se também a um mecanismo de gerenciamento no qual auxiliaria no processo de manutenção terminológica. Sabemos que a criação de uma base terminológica é uma tarefa árdua e que demanda muito tempo, corroborando com isso Foo (2012) afirma que:

> No entanto, uma terminologia deverá ser mantida, e, eventualmente, mais tempo será gasto na manutenção da terminologia comparado à criação. (FOO, 2012, p. 59, tradução nossa)[50]

Ao se tratar da manutenção e gerenciamento da terminologia, o autor nos aponta ainda quatro tarefas que merecem destaque e são de suma importância, *viz*: i) adição de novos termos *admitidos* ou *proIbidos* a um conceito (acrescentamos aqui também que no caso desta pesquisa, isso se dará não somente nos termos na língua de partida, mas também em seus correspondentes tradutórios); ii) a verificação inconsistências

[49] One issue which is highlighted by the study presented in Foo and Merkel (2010) was that in a compounding language, such as Swedish, there are more single-word terms than multi-word terms.

[50] However, a terminology must be maintained, and eventually, more time will be spent on maintaining a terminology compared to creating it.

na terminologia (tanto na busca pelos candidatos a termos, como também pelo correto posicionamento dos correspondentes tradutórios enquanto termos na língua de chegada, tomando cuidando para com sua contextualização); iii) adição de novos termos a terminologia existente (essa tarefa exigira um mecanismo mais bem elaborado, pois não somente a questão da identificação dos candidatos a termos e sua busca pelos correspondentes tradutórios, o sistema exigirá um trabalho detalhado na inclusão de itens em uma base de dados devidamente preparada para essa finalidade; iv) harmonização terminológica.

Por fim, sabemos da importância do uso da terminologia. Em algumas áreas não e possível ter um grande número de materiais disponível para a compilação de corpora de tamanhos razoáveis. Outros fatores também podem estar relacionados com a falta de recursos textuais para a criação e novos bancos terminológicos que possam auxiliar tradutores. Neste trabalho conseguiu-se amparo legal para a utilização dos textos. Com base no documento do POINTER, intitulado *Terminology Resources*, podemos perceber alguns entraves que podem ser encontrados ao se trabalhar com recursos terminológicos: entraves legais, *viz.* direitos autorais, direitos de propriedade intelectual; financeiros, *viz.* custos da terminologia, faturamento de serviços on-line; de confidencialidade/propriedade e; formatos para intercambio de terminologia.

A terminologia extraída, como também os processos descritos para o desenho base da proposta desenvolvida nessa pesquisa vai além de sua utilidade para o trabalho do tradutor, podendo ser utilizada como ferramenta também para o ensino. No ensino da tradução a disciplina de terminologia é de extrema relevância, Pym (2011) sendo uns dos pesquisadores mais reconhecidos na área do ensino da tradução, destaca o contraste entre a terminologia e a tradução:

> Terminologia versus tradução: Se uma distinção deve ser feita, vamos propor a seguinte: tradução envolve a obrigação de escolher entre mais de uma solução viável para um problema, enquanto a terminologia busca situações em que há apenas uma solução viável. (PYM, 2011, p. 93, tradução automática) [51]

[51] Terminology vs. translation: If a distinction must be made, let us propose the following: translation involves the obligation to select between more than one viable solution to a problem, whereas terminology seeks situations where there is only one viable solution.

As palavras de Pym (Ibid.) mostram claramente a grande diferença, mesmo que ambas as disciplinas façam parte uma da outra, ora, a terminologia pode ser encarada como uma subárea da tradução; ora, a tradução pode ser encarada como um dos usos, uma das áreas em que a terminologia fornece subsídios, tal distinção remete ao cotidiano do tradutor. Assim como em uma via de mão única onde o motorista tem que fazer sua escolha pela qual deve seguir e não poderá voltar atrás. O tradutor, ao se deparar com itens lexicais específicos de uma área, traz aqui novamente a definição de terminologia como sendo "um conjunto de termos que representam um sistema de conceitos de uma área em particular"[52] necessitam de esforços cognitivos árduos para a resolução de um problema de tradução que, de acordo com Pym (2011, p. 94) é uma "situação em que um elemento do texto de chegada deve ser procurado com a finalidade de corresponder de alguma forma a um elemento do texto de partida e mais de uma solução é viável".

O acesso a recursos úteis para tradutores como a recursos terminológicos sejam eles em propostas, procedimentos ou ferramentas é altamente importante, pois interfere principalmente na produtividade e qualidade dos textos traduzidos, produto recorrente do processo de tradução. A globalização já esta acontecendo e quanto mais tempo gasto durante a atividade tradutória, maiores sãos os custos a serem repassados aos produtos. Desta forma, encorajam-se tradutores na criação de suas bases terminológicas com o uso de propostas como a abordada nessa tese de doutoramento.

Embora este seja apenas mais um trabalho em meio ao tantos já realizados na área da terminologia, sabemos que essa área carece cada vez mais de recursos, principalmente quando tratamos da língua portuguesa padrão brasileiro a fim de tornar o trabalho dos especialistas da área da terminologia e, principalmente, dos tradutores mais produtivo, no entanto nunca deixando de primar pela qualidade.

Finalmente, busca-se tão somente afirmar que apesar de eventuais restrições acadêmicas – já históricas – em relação à interdisciplinaridade que se estabelece em as *ciências da linguagem* e as *ciências da computação*, não se pretende absolutamente, negar que a *pressuposição*, a *polissemia*, a *ambiguidade*, bem como noções elásticas como o *humor*, o *erótico*, o *horror*, sejam comparáveis a entidades que não se deixam aprisionar facilmente por meio de cálculos matemáticos. Nesse sentido, nos apraz fechar o texto com uma breve citação de

[52] Terminology is set of terms representing a system of concepts of a particular subject field and the discipline dealing with it.

Georges Bataille que, em seu texto intitulado *Informe*, apresentado na *Revista Documents* em 1929/30 (dezembro/janeiro), resume parte de seu debate com Carls Einstein a respeito do Surrealismo, mas que pode perfeitamente interpretado como resposta a discussão que sacudiram os ânimos de linguistas da UNICAMP no Brasil dos 1980 e que hoje ainda reascendem debates entre terminólogos, lexicógrafos e tradutores: a questão das necessidades do discurso e as estratificações e categorizações da ciência.

Informe

"*Un dictionnaire commencerait à partir du moment où il ne donnerait plus de sens mais la besogne des mots. Ainsi informe n'est pas seulement un adjectif ayant tel sens mais un terme servant à déclasser, exigeant généralement que chaque chose ait sa forme. Ce qu'il désigne n'a ses droits dans aucun sens et se fait écraser partout comme une araignée ou un ver de terre. Il faudrait, en effet, pour que les hommes académiques soient contents, que l'univers prenne forme. La philosophie entière n'a d'autre but : il s'agit de donner une redingote à ce qui est, une redingote mathématique. Par contre affirmer que l'univers ne ressemble à rien et n'est qu'informe revient à dire que l'univers est quelque chose comme une araignée ou un crachat*." (BATAILLE, G., 1896 - 1962).

Informe (disforme)

Um **dicionário** começaria a ser considerado como tal a partir do momento em que não oferecesse mais o **sentido das palavras**, mas as necessidades que temos delas. Assim, *informe* (disforme) não é somente um adjetivo tendo esse sentido, mas um termo servindo para desclassificar, exigindo que cada coisa tenha uma forma. O que essa palavra designa não tem seus direitos garantidos em nenhum sentido e pode ser destruído facilmente como um inseto ou um verme. Seria preciso, efetivamente, para que os acadêmicos fiquem satisfeitos, que o universo tome forma. A filosofia inteira não tem outro objetivo: trata-se de conceder uma roupagem àquilo que é, uma roupagem matemática. Contrariamente, afirma que o universo não se parece com nada e não é senão algo informe (disforme) equivaleria a dizer que o universo é algo como um inseto esmagado ou um cuspe lançado ao chão. (BATAILLE, G., 1896 – 1962, tradução e grifo nosso).

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.

ALMEIDA, G. M. de B. **O percurso da Terminologia de atividade prática à consolidação de uma disciplina autônoma**. TradTerm, São Paulo, v. v. 9, p. 211-222. Disponível em: <http://www.gel.org.br/estudoslinguisticos/volumes/32/htm/ mesaredo/mr004.htm>. 2003. Acesso em: 20de junho de 2011.

ALTENBERG, B.; AIJMER, K. **The English-Swedish Parallel Corpus**: *A resource for contrastive research and translation studies*. In: Mair, Christian; Marianne Hundt (Orgs.). Corpus Linguistics and Linguistic Theory. Papers from the 20th International Conference on English Language Research on Computerized Corpora (ICAME 20) Freiburg im Breisgau 1999. Amsterdam/Philadelphia,Rodopi, 2000. p. 15-33

ARCHER, J. **Internationalisation, technology and translation**. Perspectives. Studies in Translatology, 10, 2002. p. 87-117

ARNOLD, D. J.; BALLAN, L.; MEIJER, S.; LEE HUM-PHREYS, R.; SADLER, L. **Machine Translation**: *an In-troductory Guide*. Blackwells-NCC, London, 1993.

ASKEHAVE, I. **The Internet for teaching translation**. Perspectives. Studies inTranslatology 8. p. 135-143, 2000.

BAKER, M. **Corpora in translation studies**: *An overview and some suggestions for future research*. Target, Amster-dam, John Benjamins, v. 7, n. 2, 1995. p. 223-243

BAKER, M. **Corpus linguistics and translation studies***: implications and applications*. In: BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. (Ed.). Text and technology: in honour of John Sinclair. Amsterdam: John Benjamins, 1993. p. 233-250

BAKER, M. **Towards a Methodology for Investigating the Style of a Literary Translator**. Target, Amsterdam v.2 n.12, 2000. p. 241-266

BARROS, L. A. **Curso básico de Terminologia**. São Paulo: EdUSP, 2004.

BARTHES, R. **A morte do autor**. In: O rumor da língua. Trad. Mário Laranjeira. São Paulo: Ed. Brasiliense, 1988.

BARTHES, R. **Aula**. (Tradução e Posfácio Leyla Perrone-Moisés). São Paulo: 1978.

BARTHES, R. **O prazer do texto**. (Trad. J. Guinsburg) 5. ed. São Paulo: Perspectiva, 2010.

BARTHES, R. **O Rumor da Língua**. São Paulo: Martins Fontes, 2004.

BATAILLE, G. **Informe**. Documents 7. Dec. 1929/30. p. 382.

BEEBY, A. **Choosing and empirical-experimental model for investigating translation competence**. Maeve Olohan (Org.) (2000). Intercultural Faultiness – research models in Translation Studies I – Textual and cognitive aspects. Manchester: St Jerome, 2000. p. 43-55

BERBER SARDINHA, T. **Corpora eletrônicos na pesquisa em tradução**. Cadernos de Tradução (UFSC), Florianópolis, Santa Catarina, v. 9, n. 1, 2002. p. 15-60

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri, SP: Manole, 2004.

BERBER SARDINHA, T. **Padrões lexicais e colocações do português**. Apresentação no XV ENPULI, São Paulo, USP. 1999.

BERGLUND, Y. **Future in Present-day English**: *Corpus-based evidence on the rivalry of expressions*. ICAME Journal 21, 1997. p. 7-19

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus linguistics**: *investigating language structure and use*. Cambridge, Cambridge University Press, 1998.

BIOY CASARES, A. La invención de Morel. 1940.

BOURIGAULT, D. **Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases**. Proceedings of the 14th International Conference on Computational Linguistics, 1992.

BOWKER, L. **Computer-Aided Translation Technology**: *a practical introduction*. Ottawa (CA): University of Ottawa Press, 2002.

BOWKER, L.; PEARSON, J. **Working with specialized language**: *a practical guide to using corpora*. Routledge, 2002. p. 242

BYRNE, J. **The Coming of Age of Technical Translation**: *an Introduction. Journal of Specialised Translation*. V. 11, 2009. p. 2-5

CABRÉ, M. T. **La terminología: teoría, metodología, aplicaciones**. Barcelona: Antártida/Empúries, 1993.

CAMPION, M.; ELLEY, W. **An Academic Vocabulary List**. Wellington: New Zealand Council for Educational Research, 1971.

CAMPOS,C.F. **Fundamentos de Terminologia**. UFOP.1992.

CONSTABLE, P. **Character set encoding basics**. Implementing Writing Systems: An introduction. SIL International, 2001.

COULTHARD, R. J. **The application of Corpus Method-ology to Translation**: *the JPED parallel corpus and the Pediatrics comparable corpus*. Florianópolis: UFSC, 2005. 155 f. Dissertação (Mestrado em Estudos da Tradução), Programa de Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina, Florianópolis, 2005.

DAILLE, B. **Study and Implementation of Combined Techniques for Automatic Extraction of Terminology in The Balancing Act**: *Combining Symbolic and Statistical Approaches to Language*. New Mexico State University, Las Cruces, 1994.

DANIELSSON, P.; RIDINGS, D. **Practical presentation of a "vanilla" aligner**. In TELRI Workshop in alignment and exploitation of texts, February, 1997.

DELISLE, J. **Enseignement de la traduction**: *et traduction dans l'enseignement*. Ottawa, Canada: Les Presses de L'Université D'Ottawa, 1998.

DEMO, P. **Metodologia do conhecimento científico**. São Paulo: Atlas, 2000.

DIAS, C. A. **Terminologia: conceitos e aplicações**. Ciência da Informação, Brasília, v. 29, n. 1, p. 90-92, jan./abr. 2000. Disponível em:<http://www.scielo.br/pdf/ci/v29n1/v29n1a9.pdf>. Acesso em: 20-03-2013.

DIT. **Pointer Final Report**. 1996. p. 232.

DUNNING, T. **Accurate Methods for the Statistics of Surprise and Coincidence**. 1993.

EAGLES Guidelines. Expert Advisory Group on Language Engineering Standards, 1996.

FONSECA, L. C. **Compilação de corpora**: *aspectos Jurídicos. IX Encontro Nacional e III Internacional de Tradutores*. 2004.

FOO, J. **Computational Terminology**: *Exploring Bilingual and Monolingual Term Extraction*. Tese de Doutorado. Linköping: Linköping University Electronic Press. 2012.

FOO, J. e MERKEL, M. **Computer aided term bank creation and standardization**: *Building standardized term banks through automated term extraction and advanced editing tools*. In M. Thelen & F. Steurs (Orgs.), Terminology in Everyday Life (13, pp. 163–180). Terminology and Lexicography Research and Practice. John Benjamins Publishing Company. 2010.

FOO, J.; MERKEL, M. **Using machine learning to perform automatic term recognition**. In N. Bel, B. Daille & A. Vasiljevs (Orgs.), Proceedings of the Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods held in conjunction with the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010. p. 49–54.

FOUCAULT, M. **O que é um autor?** Lisboa: Veja/Passagens, 2002.

FOUCAULT, M. **Vigiar e punir**: *nascimento da prisão*. Petrópolis: Vozes, 2001.

FROMM, G. **A criação de um site de análise Terminológica e a possibilidade de treinamento de aprendizes**: *trabalhando com corpora técnicos*. In: SEMINÁRIO DO GEL, 58, 2010, 2010, São Carlos. Programação, 2010.

FROMM, G. **Corpus monolíngüe de informática para estudos terminológicos e terminográficos**. In: 52º Seminário do GEL, 2004, Campinas. 52º Seminário do GEL - Programação e resumos, 2004. p. 132-132

FROMM, G. **O processo de consulta modular**: *primeiros passos na construção de uma nova ferramenta para os tradutores*. In: VII Mini Enapol de Lexicologia, Lexicografia, Terminologia, Toponímia e Tradução, 2004, São Paulo. VII Mini Enapol de Lexicologia, Lexicografia, Terminologia, Toponímia e Tradução, 2004.

FROMM, G. **Padronização da microestrutura em vocabulários técnicos**: *a questão do público-alvo*. In: VII ENAPOL, 2004, São Paulo. Programação e Caderno de Resumos, 2004. p. 64-65.

FROMM, G. **VoTec: a construção de vocabulários eletrônicos para aprendizes de tradução**. São Paulo, 2007. Tese (Doutorado em Estudos Linguísticos e Literários em Língua Inglesa). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

GALE, W A.; CHURCH, K. W. **A Program for Aligning Sentences in Bilingual Corpora.** Computa-tional Linguistics 19 (1), 1993. p. 75-102

GALE, William A.; CHURCH, Kenneth W. **A program for aligning sentences in bilingual corpora**. In Meeting of the Association for Computational Linguistics, 1991. p. 177-184

GONÇALVES, J. L. V. R. & MACHADO, I. T. N. **Um panorama do ensino de tradução e a busca da competência do tradutor**. Cadernos de Tradução, Santa Catarina: UFSC, n. 17, 2006.

GRANGER, S. **A multi-contrastive approach to the use of linkwords by advanced learners of English**: *Evidence from the International Corpus of Learner English*. Paper presented at the 'Pragmatic markers in contrast' workshop organized by the Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, Brussels, 2003. p. 22-23

GRANGER, S. **The International Corpus of Learner English**: *A new resource for foreign language learning and teaching and second language acquisition research*. TESOL Quarterly, special issue on corpus linguistics, 2003.

GRIES, S. T.; STEFANOWITSCH, A. **Extending collo-structional analysis**: *A corpus-based perspectives on 'alter-nations'*. International Journal of Corpus Linguistics 9.1, 2004. p. 97-129

HATCHER, **A. J. An introduction to the analysis** of English noun compounds. In World, 16, 1960. p. 356-373

HATIM, B. **Teaching and researching translation**. London/New York: Longman, 2001.

HIEMSTRA, D. **Multilingual domain modeling in twenty-one**: *automatic creation of a bi-directional lexicon from a parallel corpus*. Technical report, University of Twente, Parlevink Group, 1998.

HOARD, J. E. **Language understanding and the emerging alignment of linguistics and natural language processing**. In: Using Computers in Linguistics: A Practical Guide. J. Lawler, H. Aristar Dry (Orgs.). Routledge, London. 1998. p. 197-230.

HOLMES, J. S. **The Name and Nature of Translation Studies**. In: James S. Holmes, Translated! Papers on Literary Translation and Translation Studies. Amsterdam: Rodopi, 1972/1988. p. 67-80.

HOLMES, J. **The Name and Nature of Translation Studies**. In: Venuti, L. (ed.): The Translation Studies Reader. London & New York: Routledge, 1972/2000. p. 172-185

HUNSTON, S. **Corpora and applied linguistics**. Cam-bridge: Cambridge University Press, 2002.

HUTCHINS, W. J.; SOMMERS, H. L. **An Introduction to Machine Translation**. Academic Press, 1992.

JACQUEMIN, C. **Spotting and discovering terms through natural language processing**. Cambridge: MIT Press, 2001.

JAKOBSON, R. **On Linguistic Aspects of Translation**. 1959. p. 113-119. In Translation Studies Reader. L. Venuti. New York: Routledge, 2000.

KAGEURA, Kyo; DAILLE, Béatrice; NAKAGAWA, Hiroshi; CHIEN, Lee-Feng. **Recent trends in computational terminology. Terminology**, John Benjamins Publishing Company, V. 10, N. 1, 2004 , pp. 1-21(21)

KENNEDY, G. **'Between' and 'through'**: *The company they keep and the functions they serve*. In: K. AIJMER & B. ALTENBERG (Orgs.). English Corpus Linguistics – Studies in honour of Jan Svartvik. London/New York: Longman, 1991.

KILGARRIFF, A. **Business Models for Dictionaries and NLP**. International Journal of Lexicography 13, 2000. 107–118.

KRIEGER, M. da G. **Terminologia**: *uma entrevista com Maria da Graça Krieger*. ReVEL, v. 9, n.17, 2011. Disponível em <http://www.revel.inf.br/files/entrevistas/revel_17_entrevista_maria_graca_krieger.pdf>. Acesso em 10 de agosto de 2013.

LAMBERT, J.; VAN GORP, H. **On Describing translations**. In: HERMANS, Theo (Ed.) The manipulation of literature. Studies in literary translation. London/Sydney: Croom Helm, 1985. p. 42-53.

LAVIOSA, S. **The English Comparable Corpus**: *A Re-source and a Methodology*. IN: BOWKER L.; CRONIN M.; KENNY D. e PEARSON, J. (Orgs.) Unity in Diversity: Current Trends in Translation Studies. Manchester: St. Jerome Publishing, 1998.

LEECH, G. **Corpora and theories of linguistic performance**. In: J. SVARTVIK (org.). Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, Berlin, New York: De Gruyter, 1992.

MARIAN, J. **O uso de corpora como ferramenta de apoio para tradução**: *uma análise das co-ocorrências do item lexical "hearing"*. Florianópolis: UFSC, 2010. 86 f. Dissertação (Mestrado em Estudos da Tradução), Programa de Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina, Florianópolis, 2010.

MATUDA, S. **Fraseologia no futebol**: *um estudo bilíngüe baseado em corpus*. 2010.

MCENERY, T.; WILSON, A. **Corpus Linguistics**. Edinburgh: Edinburgh University Press, 1996.

MCLUHAN, M. **O meio é a mensagem**. Ed. Record. Tradução: Ivan Pedro de Martins, 1969.

MONTGOMERY, S. L. **Science in Translation**. Chicago, London: The University of Chicago Press, 2000.

OCH, F. J. **Giza++**: *Training of statistical translation models*. 2000.

PARTINGTON, A. **Patterns and Meanings – Using Corpora for English Language Research and Teaching**. Studies in Corpus Linguistics 2. Amsterdam/Philadelphia: John Benjamins, 1998.

PEARSON, K. **On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably sup-posed to have arisen from random sampling**. Philosophi-cal Magazine, Series 5 50 (302), 1900. p. 157–175

PORTOLAN, A. C. **Uma contribuição para a elaboração de um glossário bilíngue na área de pediatria com base em Linguística de Córpus**: UFSC, 2011. 151 f. Dissertação (Mestrado em Estudos da Tradução), Programa de Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina, Florianópolis, 2011.

PRANINSKAS, J. **American University Word List**. Lon-don: Longman, 1972.

PYM, A. **Translation research terms: a tentative glossary for moments of perplexity and dispute**. In: From Translation Research Projects 3. Anthony Pym (Org.).Tarragona: Intercultural Studies Group, 2011. pp. 75-110. Disponível em: <http://isg.urv.es/publicity/isg/publications/trp_3_2011/index.htm> Acesso em 10 de agosto de 2013.

QIN G., NGUYEN B., STEPHAN, V. **A Semi-supervised Word Alignment Algorithm with Partial Manual Alignments**. ACL 2010 Joint fifth workshop on statistical machine translation and metrics MATR. 2010. p. 1-10.

RAYSON, P.; GARSIDE, R. **Comparing corpora using frequency profiling**. In proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), Hong Kong, 2000. p. 1-6

RAYSON, P.; LEECH, G.; HODGES, M. **Social differentiation in the use of English vocabulary**: *some analyses of the conversational component of the British National Corpus*. International Journal of Corpus Linguistics. Volume 2, number 1. John Benjamins, Amsterdam/Philadelphia, 1997. p. 133-152

SANTOS, A. G. **Colocações adverbiais em inglês para negócios**: *uma proposta à luz da Lingüística de Corpus*. 2010.

SCHRUZ, F.D. **Glossary of Terms Used in Database Searching**. http://library.iusb.edu/instruction/helpguide/handouts/DatabaseSearching.shtml, 2012.

SAUSSURE, F. de. **Curso de lingüística geral**. Trad de A. Chelini , José P. Paes e I. Blikstein. São Paulo: Cultrix; USP, 1969.

SCOTT, M. **WordSmith Tools version 2**, Oxford: Oxford University Press, 1997.

SCOTT, M. **WordSmith Tools version 3**, Oxford: Oxford University Press, 1999.

SCOTT, M. **WordSmith Tools version 4**, Oxford: Oxford University Press, 2004.

SCOTT, M. **WordSmith Tools version 5**, Liverpool: Lexical Analysis Software, 2008.

SCOTT, M. **WordSmith Tools**, Oxford: Oxford University Press, 1996.

SILVA, J. M., LUZ, J. A. M. **Enfatizando a Importância do Trabalho com Terminologia Técnica nas Escolas de Engenharia**. COBENGE 2004. 2004. Disponível em: <http://www.abenge.org.br/CobengeAnteriores/2004/artigos/07_283.pdf >. Acesso em: 01 de junho de 2013.

SILVA, N. A. **Análise da tradução do item lexical eviden-ce para o português com base em um corpus jurídico**. Florianópolis: UFSC, 2008. 114 f. Dissertação (Mestrado em Estudos da Tradução), Programa de Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina, Florianópolis, 2008.
SIMÕES, A. M.; ALMEIDA, J. J. **NATools** : *a statistical word aligner workbench*. In CONGRESO DE LA SEPLN, 19, Madrid, 2003.

SINCLAIR, J. **Corpus, concordance, collocation**. Oxford: Oxford University Press, 1991.

SINCLAIR, J. M. **Preliminary Recommendations on Corpus Typology**. EAGLES. 1996. Disponível em: http://www.ilc.cnr.it/EAGLES/corpustyp/node5.html. Acessado em 16 de agosto de 2013.

SIQUEIRA NOBILE, M. G. C. **Tradução e Lexicografia Jurídica no Brasil – Análise de dois Dicionários Jurídicos Português-Inglês brasileiros, considerando as peculiarida-des e os condicionantes culturais dos diferentes sistemas e linguagens jurídicas**. Florianópolis: UFSC, 2008. 158 f. Dissertação (Mestrado em Estudos da Tradução), Progra-ma de Pós-Graduação em Estudos da Tradução, Universi-dade Federal de Santa Catarina, Florianópolis, 2008.

SOUZA, M. A. O **Dicionário de Hebraico Bíblico de Brown, Driver e Briggs Como Modelo de um Sistema Lexical Bilíngüe**: *um Estudo da Lexicografia Hebraica Bíblica Moderna*. Florianópolis: UFSC, 2008. 197 f. Dis-sertação (Mestrado em Estudos da Tradução), Programa de

Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina, Florianópolis, 2008.

STUBBS, M. **Collocations and semantic profiles**: *on the cause of trouble with quantitative studies*. Functions of Language, Amsterdam, John Benjamins. 1995.

STUBBS, M. **Words and phrases**: *corpus studies of lexical semantics*. Oxford: Blackwell, 2001.

TAGNIN, S. **A multilingual learner corpus in Brazil**. Lancaster: Corpus Linguistics - Learner Corpus Workshop, 2003.

TAGNIN, S.; FROMM, G. **CoMAprend – a experiência da construção de um corpus de aprendizes para estudos**. 2010.

TEBEAUX, E. **The Emergence of a Tradition**: *Technical Writing in the English Renaissance* 1475-1640. New York: Baywood Publishing Company, 1997.

TEIXEIRA, E. D. **Tradução culinária e ensino**: *um exemplo de metodologia de avaliação utilizando etiquetagem e o WordSmith Tools*. 2010.

TEIXEIRA, E. D. **Tradução e Terminologia plurilíngüe - a Língüística de Corpus como proposta de aproximação**. 2005.
TOGNINI-BONELLI, E. **Corpus Linguistics at Work**. Amsterdam: John Benjamins, 2001.

TYMOCZKO, M. Computerized Corpora and the Future of translation studies: *The corpus based approach*. Ed. Sara Laviosa. Meta: Journal des Traducteurs 43:4, 1998.

VENUTI, L. **The Translator's Invisibility.** A History of Translation. Londres/New York: Routledge, 1995.

VILLAS BOAS, P. P. **Análise das correspondências de tradução inglês-português para substantivos e adjetivos compostos hifenizados da língua inglesa**: *uma abordagem de base em corpus*.

Florianópolis: UFSC, 2009. 86 f. Dissertação (Mestrado em Estudos da Tradução), Programa de Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina, Florianópolis, 2009.

WILKS, S. S. **The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses**. The Annals of Mathematical Statistics 9, 1938. p. 60–62

WILLIAMS, J; CHESTERMAN, A. **The Map**: *A Begin-ner's Guide to Doing Research in Translation Studies*. Manchester: Saint Jerome Publishing, 2002.

WÜSTER, E. **Introducción a la Teoría General de la Terminología y a la Lexicografía Terminológica**. Barcelona, Institut Univertari de Linguística Aplicada/Universitat Pompeu Fabra, 1998.

YATES, F. **Contingency table involving small numbers and the X2 test**. Supplement to the Journal of the Royal Statistical Society 1(2), 1934. p. 217–235

Yuste Frías, J. **Doblaje y paratraducción**. In Montero Domínguez, X. [Org.] Tradución para a dobraxe en Galicia, País Vasco e Cataluña. Experiencias investigadoras e profesionais, Vigo: Servizo de Publicacións da Universidade de Vigo, col. T&P n.º 4., ISBN: 84-8158-520-9. 2010. p. 25-29.

ZANETTIN, F. **Parallel Corpora in Translation Studies**: *Issues in Corpus Design and Analysis*. In: Olohan (2000), 2000. p. 105-118.