UNIVERSIDADE FEDERAL DE MINAS GERAIS

FACULDADE DE LETRAS

Rodrigo Araújo e Castro

DESENVOLVIMENTO, IMPLEMENTAÇÃO E TESTE DE FERRAMENTAS INTEGRADAS PARA ANÁLISE TEXTUAL E TRATAMENTO ESTATÍSTICO DE DADOS EM PESQUISAS LINGUÍSTICAS

Belo Horizonte

2016

Rodrigo Araújo e Castro

DESENVOLVIMENTO, IMPLEMENTAÇÃO E TESTE DE FERRAMENTAS INTEGRADAS PARA ANÁLISE TEXTUAL E TRATAMENTO ESTATÍSTICO DE DADOS EM PESQUISAS LINGUÍSTICAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Estudos Linguísticos da Faculdade de Letras da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Linguística Aplicada.

Área de concentração: Linguística Aplicada

Linha de pesquisa: Estudos da Tradução – 3B

Orientadora: Profª. Drª. Adriana Silvina Pagano

Co-orientadora: Profª. Drª. Ilka Afonso Reis

Belo Horizonte

Faculdade de Letras da Universidade Federal de Minas Gerais

2016

Ficha catalográfica elaborada pelos Bibliotecários da Biblioteca FALE/UFMG

Castro, Rodrigo Araújo e.

C355d Desenvolvimento, implementação e teste de ferramentas integradas para análise textual e tratamento estatístico de dados em pesquisas linguísticas [manuscrito] / Rodrigo Araújo e Castro. – 2016.

120 f., enc.: il (color)., tbs, grafs, (p&b)

Orientadora: Adriana Silvina Pagano.

Coorientadora: Ilka Afonso Reis

Área de concentração: Linguística Aplicada.

Linha de Pesquisa: Estudos da Tradução – 3B.

Dissertação (Mestrado) – Universidade Federal de Minas Gerais, Faculdade de Letras.

Bibliografia: f. 102-107.

Anexos: 108-120.

1. Tradução e interpretação – Teses. 2. Linguística aplicada – Teses. 3. Línguística – Processamento de dados – Teses. 4. Linguística de corpus – Teses. I. Pagano, Adriana Silvina. II. Reis, Ilka Afonso. III. Universidade Federal de Minas Gerais. Faculdade de Letras. IV. Título.

CDD: 418.02

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

FOLHA DE APROVAÇÃO

Desenvolvimento, implementação e teste de ferramentas integradas para análise textual e tratamento estatístico de dados em pesquisas linguísticas

RODRIGO ARAUJO E CASTRO

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGUÍSTICOS, como requisito para obtenção do grau de Mestre em ESTUDOS LINGUÍSTICOS, área de concentração LINGUÍSTICA APLICADA, linha de pesquisa Estudos da Tradução.

Aprovada em 15 de fevereiro de 2016, pela banca constituída pelos membros:

Prof(a). Adriana Silvina Pagano - Orientadora
UFMG

Prof(a). Ilka Afonso Reis - Coorientadora
UFMG

Prof(a). Igor Antônio Lourenço da Silva
UFU

Prof(a). Leonardo Pereira Nunes
UFMG

Belo Horizonte, 15 de fevereiro de 2016.

AGRADECIMENTOS

Em primeiro lugar, agradeço à minha orientadora Adriana Silvina Pagano e à minha co-orientadora Ilka Afonso Reis, que me apoiaram, fizeram críticas e, principalmente, me auxiliaram nessa longa caminhada até me tornar Mestre.

Agradeço também aos amigos do LETRA, em especial Adriana Alves, Aline, André, Arthur, Flávia, Kícila, Juliana e Luana que me acompanharam nessa caminhada e me ajudaram cada um à sua maneira.

Agradeço também à minha família, especialmente minha mãe e meu avô, que sempre me apoiaram nas minhas conquistas e estiveram próximos quando eu precisasse.

Não posso deixar de citar minha namorada, Cristiane, que conheci apenas no começo de 2015 e que, apesar de sermos bem diferentes um do outro, já mudou minha vida para melhor e continua fazendo-o a cada passo da nossa caminhada pela vida.

Por fim, para não deixar de mencionar o apoio de outras pessoas importantes, agradeço a meu amigo Paulo e a outras grandes presenças, como Giacomo, Francieli e Luís Guilherme, que também deixaram sua contribuição para este valioso trabalho.

‘‘Não vim até aqui
Pra desistir agora
Entendo você
Se você quiser ir embora
Não vai ser a primeira vez
Nas últimas 24 horas
Mas eu não vim até aqui
Pra desistir agora

Minhas raízes estão no ar
Minha casa é qualquer lugar
Se depender de mim
Eu vou até o fim
Voando sem instrumentos
Ao sabor do vento
Se depender de mim
Eu vou até o fim’’

Trecho de “Até o fim” (Engenheiros do Havaí)
Compositor: Humberto Gessinger

RESUMO

Esta dissertação apresenta um estudo de desenvolvimento, implementação e teste de um conjunto de ferramentas de preparação e análise de dados estruturados (em planilhas) e não estruturados textuais, utilizando-se *scripts* elaborados no *software* estatístico e ambiente computacional *R*. Contribuindo para os Estudos da Tradução, no escopo da Linguística com potencial de aplicação (HALLIDAY, 1985), desenhada no marco teórico da Linguística Sistêmico-Funcional (HALLIDAY; MATTHIESSEN, 2014), e utilizando subsídios da Linguística de *Corpus*, da Mineração de dados e de textos, da Estatística Descritiva e de técnicas multivariadas de análise, foram desenvolvidos e testados *scripts* em dados provenientes de um estudo experimental realizado no Laboratório Experimental de Tradução, da Faculdade de Letras da Universidade Federal de Minas Gerais, com quatro pesquisadores do Centro de Desenvolvimento de Tecnologia Nuclear e quatro tradutores profissionais. Os dados selecionados consistiram em (i) dados sociodemográficos e informações fornecidas pelos sujeitos do experimento, como hábitos de leitura e conhecimentos linguísticos na L1 e L2 constituindo dados do tipo estruturado; e (ii) dados do tipo não estruturado extraídos de protocolos verbais (livres e guiados) da tarefa realizada pelos sujeitos. A preparação dos dados estruturados foi feita no *R*, bem como sua análise, que enfocou a sumarização de dados desses sujeitos, triangulados com o agrupamento feito por técnica de análise multivariada. A preparação dos dados não estruturados foi feita com o editor de texto *Notepad++* e *scripts* do ambiente *R*, também utilizados para a análise dos dados, enfocando-se os pronomes "eu" e "a gente" e sua co-ocorrência com verbos, enquanto realizações das categorias do sistema de TRANSITIVIDADE PARTICIPANTE e PROCESSO, passíveis de serem analisadas sob a perspectiva da metarreflexão dos sujeitos do experimento sobre a tarefa executada. A análise dos dados estruturados permitiu agrupar os sujeitos dos experimentos e obter dendrogramas com base nas planilhas de dados. A análise dos dados não estruturados permitiu a obtenção de: lista de frequência, nuvem de palavras, linhas de concordância e lista de colocados. Os resultados do estudo de implementação evidenciaram os sujeitos mais similares dentro de cada grupo e na amostra como um todo, assim como o fato de os verbos em co-ocorrência com os pronomes examinados nos protocolos serem aqueles que realizavam PROCESSOS materiais e relacionais (relacionados à representação de atividades de fazer e atribuir), seguidos dos mentais (incluindo instâncias de metáforas interpessoais), os quais, segundo Magalhães e Alves (2006), sugerem de forma "mais deliberada" a metarreflexão dos sujeitos.

Palavras-chave: Linguística com potencial de aplicação; Estudos da Tradução; Linguística Sistêmico-Funcional; Mineração de dados; Mineração de textos.

ABSTRACT

This thesis reports on a study aimed at developing, applying and testing a set of tools designed for the pre-processing and analysis of structured (spreadsheet) and unstructured data by means of scripts written in the R software and environment. Contributing to Translation Studies, within the scope of appliable linguistics (Halliday, 1985), as conceived of by Systemic Functional Linguistics (Halliday and Matthiessen, 2014), and drawing on Corpus Linguistics, data and text mining and descriptive and multivariate statistics, scripts were written and tested on data retrieved from a study carried out at the Laboratory for Experimentation in Translation, Arts Faculty, Federal University of Minas Gerais, in which four nuclear scientists of the Center for the Development of Nuclear Energy, and four professional translators were asked to produce a translation in an experimental setting. The data set selected were (i) subjects' sociodemographic data and their answers to a questionnaire on their reading and writing habits and proficiency in L1 and L2 (structured data in spreadsheets) ; and (ii) unstructured data (text) retrieved from recall protocols carried out by subjects upon task completion. Structured data were pre-processed in the R environment through designed scripts. The focus of the analysis was summarizing the subjects' data, which were triangulated with the clustering results generated through the multivariate analysis technique. Unstructured data were pre-processed in the Notepad++ text editor and through designed scripts in order to analyze the pronouns "eu" and "a gente" and verbs co-occurring with them as realizations of PARTICIPANT and PROCESS categories within the TRANSITIVITY system ascribable to instances of subjects' metareflection on their task. Structured data analysis allowed for clustering subjects and obtaining dendrograms. Unstructured data analysis generated frequency lists, word clouds, Keywords in Context and lists of collocates. The results of the implementation study showed which subjects were more similar in each group and in the sample as a whole. They also showed that the most frequent verbs co-occuring with the selected pronouns were those realizing material and relational PROCESSES (associated to subjects representation of their task as doing and attributing activities), followed by mental PROCESSES (including instances of interpersonal metaphors), which, according to Magalhães and Alves (2006), tend to relate, more deliberately, to subjects' metareflection.

Keywords: Translation Studies; Appliable Linguistics; Systemic-Functional Linguistics; Data Mining; Text Mining.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

INTRODUÇÃO

A linguagem é objeto de estudo de diversas áreas do conhecimento, tanto na grande área de Letras, como nas Ciências Exatas. Na área de Letras, na subárea da Linguística, e mais especificamente da Linguística Aplicada, a Linguística de *Corpus* se dedica ao estudo de amostras de linguagem espontânea, compiladas de acordo com critérios específicos, e pela sua natureza de campo que incorpora tecnologias computacionais, se beneficia de métodos e ferramentas desenvolvidas por outros campos, como a Linguística Computacional e a Ciência da Computação, dentre outras. Nestas últimas, a linguagem humana espontânea é chamada de "linguagem natural", em oposição à linguagem artificial ou computacional.[1]

No caso da Linguística de *Corpus*, campo disciplinar de utilização principal das ferramentas elaboradas nesta pesquisa, seu desenvolvimento se pauta pelos avanços em outros campos, à medida que novas tecnologias são criadas e novas ferramentas são feitas para atender às novas necessidades e para permitir que soluções mais eficazes sejam utilizadas para as necessidades existentes. Nesse sentido, têm sido desenvolvidos diversos *softwares*, tanto proprietários como livres. Mais recentemente, o *software* e ambiente computacional *R*[2] começou a ser utilizado, apresentando-se como um recurso inovador para a implementação de análises de *corpora* com técnicas utilizadas pelos *softwares* concordanciadores, tais como *Antconc* e *WSTools*, em um espaço integrado de preparação, análise e tratamento estatístico de *corpora*.

O tema desta pesquisa compreende o desenvolvimento de ferramentas computacionais de análise e a descrição dos resultados gerados por essas ferramentas por meio do uso da Estatística Descritiva. A pesquisa orienta-se para a elaboração de *scripts*[3] para o sistema operacional Windows[4] no *software R* que possam operacionalizar ferramentas de análise textual em um *corpus* constituído de amostras de um conjunto de transcrições de gravações de

[1] Disponível em: <http://www.oxforddictionaries.com/us/definition/american_english/natural-language>. Acesso em: 21 jan. 2016.
[2] Ao longo do texto, são feitas várias menções ao *software R*. Para informações básicas sobre ele, sugere-se conferir Zara (2013), uma vez que é necessário ter conhecimentos básicos desse *software* e tê-lo instalado corretamente para utilizar os *scripts* fornecidos nesta dissertação.
[3] *Scripts* são conjuntos de linhas de comando que representam instruções a serem seguidas pelo computador a fim de se realizar determinada tarefa.
[4] É possível que as funções e *scripts* funcionem em outros sistemas operacionais, porém algumas funções podem não funcionar devido às diferenças entre os sistemas.

protocolos verbais livres e guiados coletados com sujeitos (tradutores e pesquisadores). Esses sujeitos participaram de experimentos no Laboratório Experimental de Tradução (LETRA) e geraram esses protocolos, os quais podem ser caracterizados como um *corpus* comparável monolíngue (em português brasileiro) com 8203 itens (*tokens*) – 2820 nos textos dos tradutores e 5383 nos textos dos pesquisadores (engenheiros nucleares) – e composto de resumos acadêmicos, a partir dos quais foram elaborados textos de popularização da ciência. Foi utilizado, também, um banco de dados estruturados de informações sobre os sujeitos cuja fala foi gravada nos referidos protocolos.

Esta dissertação insere-se nos trabalhos do grupo de pesquisa "Modelagem Sistêmico-Funcional da tradução e da produção textual multilíngue", registrado na plataforma de Grupos de Pesquisa do CNPq e formado por estudantes de graduação e de pós-graduação da Faculdade de Letras da Universidade Federal de Minas Gerais (FALE-UFMG), sob coordenação da Professora Doutora Adriana Silvina Pagano. Esse grupo de pesquisa tem como objetivo realizar um estudo empírico-experimental da produção textual em tradução orientada para a modelagem sistêmico-funcional da tradução e da produção multilíngue.

O estudo objeto desta dissertação afilia-se aos Estudos da Tradução, com base em trabalhos como Baker (2000) e Malmkjaer (2004), que, apesar de considerados quantitativos (por utilizarem dados numéricos), não utilizam os métodos quantitativos aplicados nos Estudos da Tradução (utilizando também ferramentas computacionais e estatísticas), segundo Oakes e Ji (2012) e Gries (2010). Além disso, esta pesquisa também utiliza como subsídio dois conjuntos de técnicas de outra área do conhecimento, a Mineração de dados e a Mineração de textos (ou Mineração de dados não estruturados do tipo texto), no âmbito da Ciência da Computação, estudada por pesquisadores como Feinerer, Hornik e Meyer (2008). Dessa forma, utilizando o ambiente computacional *R*,[5] foram elaborados *scripts* (ver Anexos) com o objetivo de criar ferramentas de análise computacional de dados textuais e numéricos.

Esta dissertação também se inclui no escopo da *Linguística com potencial de aplicação* (HALLIDAY, 1985), desenhada no marco teórico da Linguística Sistêmico-Funcional (HALLIDAY; MATTHIESSEN, 2014), e utilizando subsídios da Linguística de *Corpus*, da Mineração de dados e de textos, da Estatística Descritiva e de técnicas multivariadas de análise.

[5] O *R* é, ao mesmo tempo, um ambiente e uma linguagem de programação livre e é amplamente utilizado em análises estatísticas e em representações gráficas.

Foram desenvolvidos *scripts*, testados em dados provenientes de um estudo experimental realizado no Laboratório Experimental de Tradução (LETRA), da FALE/UFMG, com 4 pesquisadores do CDTN (Centro de Desenvolvimento de Tecnologia Nuclear) e 4 tradutores profissionais.

Como apontado por Malmkjaer (2005), a contribuição de uma pesquisa para seu campo disciplinar pode se dar de várias formas, por exemplo, por meio da aplicação do conhecimento de uma área de estudo em um campo disciplinar. No caso desta dissertação, aplicaram-se os conhecimentos de áreas como a Ciência da Computação e da Estatística no campo disciplinar dos Estudos da Tradução (MALMKJAER, 2005, p. 20). Esta dissertação contribui para esse campo disciplinar pela proposta de ferramentas de análise de dados extraídos de textos traduzidos (produto) e dados obtidos na execução de pesquisas do processo tradutório, gerados em condições experimentais. Pode-se dizer ainda que esta dissertação contribui para os Estudos da Tradução aplicando conhecimentos de outro(s) campo(s) disciplinar(es) no estudo do fenômeno da tradução, permitindo que achados e percepções de outras áreas do conhecimento auxiliem na melhor compreensão deste fenômeno (MALMKJAER, 2005, p. 20-21).

Uma vez que a Mineração de dados lida com dados estruturados (no caso desta dissertação, estruturados em planilhas de dados) e a Mineração de textos com dados não estruturados (neste caso, do tipo texto), esses são os tipos de dados analisados, e para cada um foram desenvolvidas ferramentas. Para os dados não estruturados, foram elaborados *scripts* para gerar listas de frequência,[6] nuvem de palavras,[7] linhas de concordância[8] e listas de colocados[9] (à esquerda e à direita). Já para os dados estruturados, foram gerados *scripts* para criar dendrogramas[10] (cf. ALBUQUERQUE, 2005).

---

[6] Segundo Castro e Cecílio (2015, p.106), a lista de frequência é uma lista das palavras mais frequentes no *corpus* que pode ser filtrada para a exibição de algumas consideradas mais relevantes de acordo com a escolha do pesquisador.

[7] Segundo Castro e Cecílio (2015, p.106), a nuvem de palavras é a representação gráfica das palavras mais frequentes em um texto ou *corpus*, sendo que quanto mais frequente a palavra, maior e mais centralizada ela será.

[8] Segundo Castro e Cecílio (2015, p.106), linhas de concordância são linhas de texto que indicam as palavras que ocorrem antes e depois da palavra de busca a determinado número de posições de acordo com o interesse do pesquisador. Por exemplo, o pesquisador pode determinar que são do seu interesse as 5 palavras à esquerda e/ou à direita.

[9] Segundo Castro e Cecílio (2015, p.106), listas de colocados são listas das palavras mais frequentes à esquerda ou à direita da palavra de busca a determinado número de posições de acordo com o interesse do pesquisador, como, por exemplo, 5 posições à palavras à esquerda e/ou à direita.

[10] Um dendrograma consiste na representação gráfica do agrupamento de diversos elementos analisados, seja a partir de sua similaridade ou de sua dissimilaridade e pode ser feito por meio de diversos métodos matemáticos (como o método Ward) e utilizar como base uma das diversas distâncias matemáticas, como a distância euclidiana, ou distância entre pontos.

A partir dos resultados obtidos pelas ferramentas elaboradas (funções[11] dos *scripts*) para os dados não estruturados, enfocaram-se os processos mentais, que, segundo Magalhães e Alves (2006), apresentam de forma "mais deliberada" a metarreflexão. Para esse fim, foram analisados os pronomes "eu" e "a gente" em co-ocorrência com grupos verbais (relacionados à classe de palavra "verbo") realizando PROCESSOS,[12] os quais possuem ou não realizações da categoria PARTICIPANTE do sistema de TRANSITIVIDADE.[13] [14] Essas categorias são passíveis de serem analisadas sob a perspectiva discursiva como indicadores do posicionamento dos sujeitos dos experimentos em suas falas sobre a tarefa de tradução executada. Não foi enfocado o pronome "nós" pois, como será visto no capítulo 3, não houve ocorrência de pronomes pessoais diferentes de "eu" e "a gente" com função de sujeito dentre os 20 termos relevantes mais frequentes.

Quanto aos resultados dos dados estruturados, triangularam-se os dados sumarizados por meio de quadros e gráficos com os dendrogramas gerados a partir das planilhas de dados a fim de verificar a razão dos agrupamentos dos sujeitos nos dendrogramas com base nos padrões dos dados. Como consequência, foi possível apresentar o resultado das análises dos dados estruturados, assim como da triangulação realizada, e dos não estruturados, os quais foram interpretados à luz da Linguística Sistêmico-Funcional, tendo como unidade de análise o texto, visto que todos os dados são analisados visando a geração de conclusões sobre o texto.

Esta dissertação está dividida em Introdução, os capítulos da Revisão teórica (capítulo 1), da Metodologia (capítulo 2) e da Implementação das ferramentas de análise elaboradas (capítulo 3), seguidos das Considerações finais, das Referências Bibliográficas e dos Anexos.

---

[11] Na computação, funções, da mesma forma que um *script*, são conjuntos de instruções sequenciais dados a um programa (*software*) para realizar alguma atividade.

[12] Segundo Halliday e Matthiessen (2014), PROCESSO é a função (categoria gramatical) do sistema de TRANSITIVIDADE que opera na ordem da oração (constituída prototipicamente por PARTICIPANTE, PROCESSO e CIRCUNSTÂNCIA) e que tem como realização no português brasileiro o grupo verbal (relacionado à classe de palavra "verbo").

[13] Segundo Halliday e Matthiessen (2014), a Linguística Sistêmico-Funcional, PARTICIPANTE é a função (categoria gramatical) do sistema de TRANSITIVIDADE que opera na ordem da oração (constituída prototipicamente por PARTICIPANTE, PROCESSO e CIRCUNSTÂNCIA) e que tem como realização no português brasileiro o grupo nominal (relacionado à classe de palavra "verbo").

[14] Nesta dissertação, todos os termos em versalete são termos técnicos da Linguística Sistêmico-Funcional, a fim de evitar ambiguidades na compreensão da leitura.

Objetivos

Objetivo geral

- Desenvolver, implementar e testar um conjunto de ferramentas de preparação e análise de dados estruturados (planilhas de dados) e não estruturados (protocolos verbais) para a realização de estatísticas aplicáveis a esses dados, explorando-se a potencialidade do ambiente *R*.

Objetivos específicos

- Elaborar *scripts* do ambiente *R* para auxiliar na preparação de dados estruturados e não estruturados;
- Desenvolver uma metodologia baseada na Mineração de dados e de textos com base em ferramentas da Linguística de *Corpus* para extrair os dados de planilhas e de textos no ambiente *R*;
- Implementar e testar os *scripts* elaborados em *corpus* constituído de uma amostra de transcrições de protocolos verbais de experimentos (dados não estruturados) e em um banco de dados de informações relativas aos tradutores envolvidas nesses experimentos (dados estruturados), com o objetivo de triangular os dados gerados e as estatísticas aplicadas;
- Interpretar os resultados da implementação das ferramentas elaboradas utilizando como subsídio a Linguística Sistêmico-Funcional, mais especificamente na análise dos PROCESSOS mentais sob a perspectiva discursiva como indicadores do posicionamento dos sujeitos dos experimentos em seus protocolos verbais sobre a tarefa de tradução executada.

CAPÍTULO 1

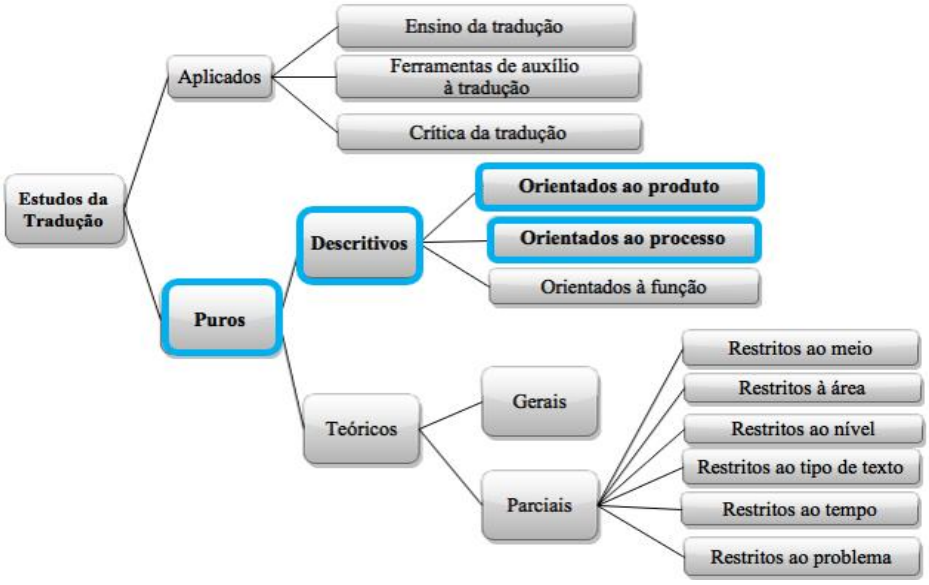
OS ESTUDOS DA TRADUÇÃO NO ESCOPO DE UMA LINGUÍSTICA COM POTENCIAL DE APLICAÇÃO

Este capítulo apresenta a revisão das principais teorias que informam esta dissertação, que se localiza no campo disciplinar dos Estudos da Tradução. No escopo desse campo disciplinar, são apresentados trabalhos que fazem uso de métodos quantitativos, os quais são contrastados com trabalhos que não utilizam tais métodos. Por fim, é apresentado o conceito de Linguística com potencial de aplicação (*Appliable linguistics*), ao qual esta dissertação se afilia utilizando como fontes de subsídios conhecimentos de áreas distintas, quais sejam, a Linguística de *Corpus*, a Mineração de dados e de textos e a Estatística Aplicada aos estudos linguísticos. Foram contemplados na revisão os trabalhos (tanto da abordagem do produto quanto do processo da tradução) que fazem uso de um *corpus*, visto que esses são os que mais contribuem com esta dissertação.

1.1 Estudos da Tradução

No escopo do campo disciplinar dos Estudos da Tradução, de acordo com o mapeamento elaborado por Toury (1995) com base em Holmes (1972/1988), esta dissertação visa contribuir para os estudos puros e descritivos, orientados para o produto e para o processo, conforme observado na Figura 1.

Figura 1 – Localização desta pesquisa no campo disciplinar Estudos da Tradução de acordo com o mapa de Holmes (1972/1988)



Fonte: Adaptado de Toury (1995, p. 16) com base em Holmes (1972/1988). Destaque meu.

No caso desta dissertação, a contribuição para esse campo disciplinar se dá pela proposta de ferramentas que permitem analisar dados extraídos de textos traduzidos (produto) e dados gerados na execução de pesquisas do processo tradutório, gerados em condições experimentais.

Na análise do produto, destaca-se a contribuição da Linguística de *Corpus*, que, como Baker (2000) ressalta, pode auxiliar os Estudos da Tradução, através, por exemplo, de ferramentas que permitem a geração de (i) contagens de itens (*tokens*), ou seja, o número de palavras dos textos utilizados em pesquisas; (ii) o tamanho médio das sentenças; (iii) a razão forma/item (*type/token ratio)*, o número de palavras não repetidas no texto dividido pelo número total de palavras no texto, utilizado para verificar a densidade lexical de textos; (iv) listas de frequência, para organizar as palavras (os termos) de um texto de acordo com sua frequência de ocorrência; (v) linhas de concordância, segmentos de texto que mostram as palavras antes e/ou depois do termo de busca, utilizadas para separar o ambiente lexical de um termo de busca por meio das palavras com esse co-ocorre; (vi) listas de colocados, listas de palavras mais frequentes à esquerda ou à direita do termo de busca; (vii) palavras-chave, para analisar se determinados termos são característicos ou não de certo texto (em comparação a um volume maior de textos do mesmo tipo); e (viii) o alinhamento de textos original e traduzido(s) ou de textos comparáveis por algum critério, como o tipo de texto ou o tradutor. Algumas dessas ferramentas, como as listas de frequência e as listas de concordância, serão mencionadas mais adiante no texto, ilustradas com base em estudos que as utilizaram.

Dentre as pesquisas que utilizam ferramentas da Linguística de *Corpus* (como as citadas anteriormente), há algumas que se afiliam ao campo dos Estudos da Tradução. Entre essas pesquisas, há aquelas que utilizam abordagens quantitativas com métodos computacionais e estatísticos (métodos quantitativos) na análise dos dados, como Ji (2012), Gries (2010) e Rybicky (2012). Há também pesquisas que não utilizam os métodos quantitativos, pois, embora utilizem dados numéricos, se limitam à interpretação desses dados sem uso de ferramentas computacionais e estatísticas, como por exemplo aquelas que enfocam a questão do estilo do tradutor e do texto traduzido, tais quais Baker (2000) e Malmkjaer (2004), respectivamente.

Baker (2000) apresenta um estudo comparativo de *corpora* de textos de dois tradutores, Peter Bush e Peter Clark, realizado com o objetivo de investigar a "impressão digital" do tradutor e possíveis diferenças entre os estilos desses tradutores. Baker (2000) investiga em que medida essas diferenças se relacionam com seus posicionamentos socioculturais, suas histórias

de vida e sua experiência como tradutores. O *corpus* utilizado no estudo foi o TEC (*Translational English Corpus*), que consiste de textos em inglês traduzidos de diversas línguas-fonte (tanto europeias quanto não europeias), distribuídos em quatro *subcorpora* de textos escritos: ficção, biografia, reportagens e revistas de bordo, e somam cerca de 10 milhões de palavras. Dentre os dados gerados a partir de ferramentas da Linguística de *Corpus*, Baker (2000) utiliza a razão forma/item, o tamanho médio das sentenças, a lista de frequência e linhas de concordância.

Malmkjaer (2004) estuda traduções de textos de Hans Christian Andersen do dinamarquês para o inglês com o objetivo de criar uma nova metodologia de análise do estilo, por ela chamada de estilística tradutória (*translational stylistics*), na qual o texto-fonte é contrastado com as traduções em busca de padrões linguísticos. O objetivo dessa metodologia é propor o estudo do estilo do texto traduzido e verificar de que maneira determinado texto traduzido foi elaborado para criar um significado possivelmente diferente do texto-fonte. Segundo essa metodologia, a tradução, sendo um texto mediado, tem pelo menos quatro características que não podem ser ignoradas: (i) o mediador possui um modo particular de interpretar o texto; (ii) a mediação tem sempre um propósito; (iii) seu propósito pode diferir do propósito do texto-fonte; (iv) o público alvo da tradução pode diferir daquele do texto-fonte. Malmkjaer (2004) aborda as escolhas e as limitações enfrentadas pelos tradutores, relacionando-as com questões interculturais e linguísticas. No entanto, dentre as ferramentas da Linguística de *Corpus* citadas, apenas é explicitado o uso de linhas de concordância e do alinhamento de textos (no caso, de exemplos), além da glosa (do dinamarquês).

Considerando as abordagens de Baker (2000) e Malmkjaer (2004), pode-se afirmar que, embora ambas as autoras contribuam para o campo disciplinar com seus estudos, suas pesquisas não utilizam métodos quantitativos. Em outras palavras, Baker (2000) e Malmkjaer (2004) chegam às conclusões de seus estudos principalmente por meio da interpretação das frequências encontradas, sem o uso de ferramentas computacionais e estatísticas que vão além da interpretação das frequências dos itens linguísticos.

Diferentemente dos trabalhos de Baker (2000) e Malmkjaer (2004), há estudos que utilizam métodos quantitativos aplicados aos Estudos da Tradução, visando demonstrar as vantagens de tais estudos sobre aqueles que não os utilizam. A seção a seguir apresenta estudos como esses, que investigam o fenômeno da tradução e que utilizam métodos quantitativos.

1.2 Métodos quantitativos aplicados aos Estudos da Tradução

Uma abordagem mais recente nos Estudos da Tradução envolve o uso de métodos quantitativos que vão além dos utilizados nos estudos mencionados anteriormente. Nesta abordagem, o foco principal está na utilização de métodos computacionais e estatísticos para sustentar os resultados e as conclusões obtidas. O principal objetivo dessa abordagem é unir a análise quantitativa à qualitativa, considerando que a análise quantitativa pode fornecer dados e gerar resultados que não poderiam ser obtidos simplesmente por meio de uma análise qualitativa. Assim, de acordo com Ji (2012, p. 70), seria possível se beneficiar dos pontos fortes de cada um desses tipos de análise (quantitativos e qualitativos) a fim de se obter novas visões sobre a natureza da tradução, sendo possível utilizar a combinação dos métodos com vários tipos de texto.

Ji (2012) argumenta que existe uma carência de metodologias quantitativas em trabalhos que utilizam a Linguística de *Corpus* e mostra como o uso de conceitos estatísticos pode corroborar as hipóteses desses estudos (relacionando texto original e textos traduzidos) e detectar tendências entre os textos. Dentre os conceitos estatísticos utilizados em seu trabalho, está o uso de dados estatisticamente significativos, de regressão linear e de teste de hipóteses. A importância de verificar se os dados são estatisticamente significativos está no fato de se verificar se as diferenças apresentadas nos dados poderiam ou não ser justificadas devido ao acaso, ou seja, se esses dados são considerados relevantes para a análise. Já a regressão linear é importante para detectar tendências nos dados analisados; no caso da comparação entre texto original e o(s) texto(s) traduzido(s), observadas as tendências entre os dados, é possível compará-las e verificar qual tradução se aproxima mais (ou menos) do original. Acerca dos testes de hipóteses, eles são utilizados para generalizar resultados de análises de uma pequena quantidade de dados para um conjunto maior de dados, levando-se em conta, inclusive, a margem de erro[15] das análises realizadas.

Ji (2012) utiliza uma abordagem quantitativa com o uso de ferramentas estatísticas nos estudos de estilística tradutória. A autora analisa duas versões chinesas da obra *Don Quijote de la Mancha*, de Cervantes – a primeira publicada em 1978 e traduzida por Yang Jiang e a outra publicada em 1995 e traduzida por Liu Jingsheng. A autora utiliza os *corpora* LCMC

[15] A margem de erro se refere ao valor do erro encontrado em medições, pois, uma vez que há sempre alguma imprecisão nas mesmas, é necessário informar a precisão utilizada (SCHEUREN, 2004).

(*Lancaster Corpus of Modern Chinese*) e UCLA (*Written Chinese Corpus*). Ji (2012) teve por objetivo comparar a relação de dependência entre os textos do *corpus* (textos fontes e textos alvos) com base nas variáveis textuais analisadas, ou seja, determinar como essas variáveis descrevem os textos. Para esse fim, fez regressões lineares utilizando dados organizados em colunas e, ao analisar os resultados das regressões, levou em conta na interpretação os resultados dos testes de hipóteses feito pelo próprio procedimento computacional. Verificou também quais variáveis foram estatisticamente significativas – ou seja, que relações intertextuais deveriam ser consideradas relevantes para a análise.

Ji (2012) ressalta a importância de *softwares* de Mineração de textos para gerar diversos tipos de informações estatísticas relacionadas à *corpora*, como a frequência absoluta (contagens de ocorrências de categorias) e a razão forma/item (*type/token ratio*) (JI, 2012, p. 54). Por exemplo, em Ji (2012) o foco do estudo foram os arcaísmos presentes nas traduções por ela examinadas, pois, segundo explica, esses são

> um mecanismo retórico central no original em espanhol, visto que é utilizado com muita frequência no discurso do protagonista, seja em monólogos ou diálogos e também possui um papel crucial na caracterização da fascinação de Dom Quixote pelo medievalismo e pelos cavaleiros – o que, por sua vez, é alvo constante de críticas e de infindável humor no livro.[16] (JI, 2012, p. 59).

Outros autores que adotam métodos quantitativos e utilizam ferramentas da Linguística de *Corpus* são Gries (2010) e Rybicki (2012), os quais se utilizam da estatística para dar sustentação aos dados, por meio do uso do *software* e ambiente computacional *R*.

Gries (2010) afirma que a Linguística de *Corpus* é considerada um campo de estudo "distributivo" (*distributional*); em outras palavras, ela enfoca dados linguísticos de distribuição de frequência (relativa à distribuição numérica de categorias), seja pela ocorrência de itens linguísticos ou pela sua co-ocorrência (GRIES, 2010, p. 268). Por isso, busca mostrar como medidas estatísticas, como, por exemplo, medições de frequências (normalizadas ou não), podem ser utilizadas para analisar: (i) a ocorrência de itens linguísticos, utilizando-se, por exemplo, listas de frequência e; (ii) a co-ocorrência de itens linguísticos, por exemplo por meio de comparações estatísticas utilizando-se cálculos. Na análise de Gries (2010) também são utilizados gráficos e dendrogramas. No caso dos gráficos, o objetivo é não apenas apresentar

[16] "[Archaism is] a central rhetorical device in the Castilian original. It is abundant in the protagonist's speech, either in the form of monologue or dialogue with other fictional characters in the novel. It plays an essential role in the characterization of Don Quijote's enchantment with medieval chivalry and knighthood, which in turn is often the target of Cervantes's merciless criticism and the endless source of humour of the book".

os resultados obtidos, mas também auxiliar na comparação entre eles. No caso dos dendrogramas, o objetivo é o agrupamento dos dados utilizando-se como critério as regularidades neles presentes. Em termos de contribuição para a Linguística de *Corpus*, além dos aspectos já mencionados, são citados diversos exemplos de aplicação da estatística em questões ligadas à Linguística de *Corpus*, como o uso de listas de frequência (incluindo frequências absolutas e relativas) e de listas de colocados.

É importante destacar que, embora os métodos quantitativos aplicados a pesquisas tenham em comum o fato de se basearem em observações de frequência de itens linguísticos (GRIES, 2014, p. 36), "apesar do uso de *corpus* e de metodologias baseadas em *corpus* nos Estudos da Tradução, há uma escassez de descrições sistemáticas de métodos quantitativos que podem ser aplicados aos Estudos da Tradução" (JI, 2012, p. 53).[17] Entretanto, deve-se ressaltar que, em geral, não se pode classificar uma pesquisa como puramente qualitativa ou quantitativa (GRIES, 2014, p. 29), embora seja possível separar aquelas que utilizam métodos estatísticos e computacionais das que não os utilizam. Dessa forma, ao se compararem estudos do Estilo (BAKER, 2000; MALMKJAER, 2004), que não utilizam métodos quantitativos, com estudos que os adotam, como Rybicki (2012), é possível observar a similaridade entre os dois tipos, conforme explicitado por Gries (2014), que afirma que

> mesmo abordagens altamente qualitativas da Linguística de *Corpus* são, em última instância, baseadas nas observações de frequências em particular (zero ou mais vezes), o que exige uma análise quantitativa (que também requer interpretação).[18] (GRIES, 2014, p. 29).

Assim, tendo como base a natureza quali-quantitativa das pesquisas da Linguística de *Corpus* e o desenvolvimento contínuo desse campo de estudo, nos últimos 50 anos as metodologias baseadas em *corpora* têm se desenvolvido de forma rápida e ampla como uma "nova" metodologia na linguística – em contraste com o uso da intuição do linguista[19] (GRIES,

[17] Minha tradução de: "Despite the increasing use of corpus material and corpus methodologies in translation studies, there is a lack of systematic descriptions of quantitative methods that may be used for corpus translation studies."

[18] "even very qualitative approaches in corpus linguistics are ultimately based on observing things with particular frequencies (0 or more times), which calls for quantitative analysis (and, of course, quantitative analysis requires interpretation)".

[19] "Over the last fifty or so years, corpus-based methods have developed into one of the most rapidly growing and most widespread 'new' methodology in linguistics. Instead of relying on intuitions of what can or cannot be said, linguists are now turning more and more to corpus data to see what is or is not said."

2014, p. 29) – a partir de uma diversificação dos usos de *corpora* nos estudos linguísticos por meio da utilização de novas ferramentas. Dessa forma,

> a motivação para ir além de meras frequências de *tokens* ou probabilidades condicionais visa separar o joio (o fato de que[, por exemplo,] nomes [em inglês] co-ocorrem com alta frequência) do trigo (dados linguisticamente relevantes de co-ocorrência).[20] (GRIES, 2014, p. 29).

Por essa razão, com base na natureza quali-quantitativa de estudos como Gries (2010, 2014) e Rybicki (2012), pode-se afirmar que pesquisas dos Estudos da Tradução que utilizam métodos quantitativos também fazem uso de outras medidas além da contagem de frequências. A razão para essa nova abordagem nos estudos linguísticos é o fato de que as frequências apenas apontam tendências já presentes nos dados, não permitindo a generalização dos resultados além das categorias linguísticas estudadas. No entanto, com o uso de medidas como o alcance (*range*[21]), pode-se obter resultados mais relevantes e um poder preditivo significativamente maior do que com o uso da frequência (GRIES, 2014, p. 29).[22]

Rybicki (2012) desenvolve estudos de estilometria, definida como "o estudo de recursos mensuráveis do estilo (literário), como comprimento de sentença, riqueza de vocabulário, frequências (de palavras, de comprimentos de palavras, formas das palavras etc.)" (RYBICKI, 2012).[23]

Na estilometria, os textos de um *corpus* são utilizados de duas formas diferentes: primeiramente é utilizado um conjunto teste de textos a fim de se produzir uma série de regras – um classificador (*classifier*) para discriminar traços característicos de cada autor – e então, em outro conjunto de textos, é utilizado esse classificador para separar amostras de textos de acordo com os traços analisados. A análise computacional, que se utiliza também da estatística para fundamentar seus procedimentos, é feita por meio de medidas de frequência das palavras/termos mais frequentes e também de distâncias entre termos (obtidas

[20] "The motivation to go beyond mere token frequencies or conditional probabilities is to separate the wheat (linguistically revealing co-occurrence data) from the chaff (the fact that nouns co-occur with the a lot."
[21] Alcance (*range*) é definido como um tipo de medida de dispersão (ou seja, da variabilidade dos dados) que equivale à diferença entre o valor máximo e o mínimo dentro de determinada amostra de dados.
[22] "even a dispersion measure as crude as range can have significant predictive power above and beyond frequency".
[23] "Stylometry, or the study of measurable features of (literary) style, such as sentence length, vocabulary richness and various frequencies (of words, word lengths, word forms, etc.)".

computacionalmente por meio de cálculos matemáticos com base, por exemplo, em pontos de um plano cartesiano). São utilizadas também medidas de dispersão (medida da variabilidade dos dados, geralmente em relação a medidas como média ou mediana) e de análise multivariada (análise estatística utilizando múltiplas variáveis como, por exemplo, dendrogramas), dentre outras, de acordo com o método de análise utilizado. Rybicki (2012) destaca o papel da estatística, tanto durante a análise quanto na validação dos dados obtidos, assim como no fato de os métodos utilizados e as fórmulas matemáticas destes serem apresentados com ênfase no decorrer do texto.

Nesta dissertação, o uso desses procedimentos computacionais e estatísticos caracteriza as pesquisas como utilizando métodos quantitativos. Há pesquisas nos Estudos da Tradução que utilizam métodos quantitativos em suas investigações e que adotam como teoria linguística a Linguística Sistêmico-Funcional (HALLIDAY; MATTHIESSEN, 2014). A Linguística Sistêmico-Funcional utiliza, em sua metodologia, comparações entre as frequências de categorias atribuídas a ocorrências de itens linguísticos extraídas de, por exemplo, diferentes tipos de texto e também o mapeamento das probabilidades de ocorrência dessas categorias. Para isso, é utilizado um conjunto de categorias retiradas de textos para mapear as frequências e as probabilidades de cada categoria, uma vez que "cada instância de um texto falado ou escrito interfere nas probabilidades gerais do sistema – de natureza probabilística" (HALLIDAY, 2005, p. 91).[24] Quanto maior o tamanho do *corpus* (em número de palavras) utilizado, mais precisa é considerada a descrição do sistema (HALLIDAY, 2005, p. 82). De acordo com Halliday (2005), os sistemas linguísticos têm natureza probabilística, isto é, podem ser descritos por meio de frequências e probabilidades associadas a cada categoria.

Um estudo desse tipo é Halliday (2005), que relata a investigação da probabilidade de ocorrência de orações positivas e negativas pelo ponto de vista da Linguística Sistêmico-Funcional. Essa pesquisa partiu de estudos anteriores de Halliday sobre a gramática da língua chinesa (HALLIDAY, 1956; 1959), da qual o autor depreende a hipótese de que sistemas gramaticais binários (com apenas duas opções) se dividem em dois tipos: aqueles com probabilidade igual (simétrica) de ocorrência, com aproximadamente 50% para cada opção, e

[24] "Every instance of a text that is spoken or written in English perturbs the overall probabilities of the system, to an infinitesimal extent (whether it has been recorded in a corpus or not! — hence the function of the corpus as a sample). To say this is to treat the system as inherently probabilistic."

aqueles com probabilidade de ocorrência diferente (assimétrica), com aproximadamente 90% para uma opção e 10% para a outra. Halliday (2005, p. 81) se voltou ao estudo da gramática da língua inglesa e analisou amostras de 2000 ocorrências (orações) para cada sistema de interesse. Suas conclusões corroboraram a hipótese criada com base no chinês sobre os dois tipos de sistemas (os de probabilidades simétricas e assimétricas). Pode-se citar, como exemplo do segundo tipo de sistema, o sistema de POLARIDADE, cujas probabilidades aproximadas de ocorrência são: positivo (90%) e negativo (10%).

Outro estudo que se utiliza de métodos quantitativos em suas investigações e que adotam como teoria linguística a Linguística Sistêmico-Funcional é Halliday e Webster (2014), que utiliza métodos quantitativos aplicados para investigar os "discursos de abertura" (*commencement speech*) da *Stanford University* proferidos por Steve Jobs em 2005 e Susan Rice em 2010. Os autores observaram as suas distribuições dos tipos de PROCESSO nos discursos. Em seguida, geraram uma nuvem de palavras, que representa graficamente a lista de frequência das palavras (ou termos) dando destaque (em fonte maior) àquelas de maior frequência. O uso deste recurso, analogamente aos métodos quantitativos, tem como fim a visualização das palavras de maior frequência, as quais, nesse caso, são consideradas como de maior relevância, como será apresentado no capítulo 3.

No âmbito dos Estudos da Tradução, a Linguística Sistêmico-Funcional é uma das teorias linguísticas utilizadas nos estudos descritivos voltados ao processo e ao produto, o que demonstra haver um conjunto de trabalhos nesse campo disciplinar informados por essa teoria linguística, embora nem todos os estudos sistêmico-funcionais se afiliem ao campo disciplinar. Em contrapartida, há estudos de base sistêmico-funcional com potencial de implementação nesse campo disciplinar. Pode-se citar como exemplo Figueredo, Pagano e Ferreguetti (2014), no qual, tendo como objeto de estudo orações relacionais de diversos registros que foram retiradas de um *corpus* paralelo bilíngue inglês-português brasileiro, investigaram-se a equivalência textual dos PROCESSOS relacionais e os padrões linguísticos poderiam ser observados a partir do seu uso. Como resultado, verificou-se, em todos os tipos de textos estudados (artigo acadêmico, discurso político, divulgação científica, ficção, manual de instrução, propaganda turística, resenha e *website* educacional), a equivalência de significados relacionais em inglês em relação aos significados relacionais no português brasileiro.

Assim como será apresentado na Figura 2, na próxima seção, pode-se afirmar que a Linguística Sistêmico-Funcional utiliza subsídios de outros campos de estudo e propõe a Linguística com potencial de aplicação, apresentada a seguir.

1.3 Linguística com potencial de aplicação

Appliable Linguistics (HALLIDAY, 1985), ou, em português, *linguística com potencial de aplicação* (PAGANO; FIGUEREDO; LUKIN, 2014), pode ser definida como: "um tipo de linguística em que a teoria é elaborada especificamente com o objetivo de ser aplicada para resolver problemas das pessoas ao redor do mundo, envolvendo tanto reflexão como ação"[25] (MATTHIESSEN, 2012, p. 436).

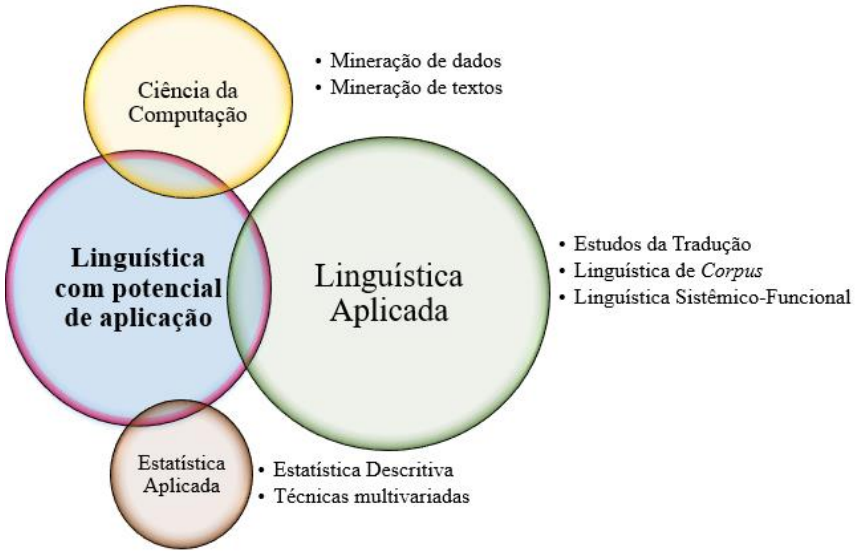
Este conceito de "linguística com potencial de aplicação" foi desenvolvido por Halliday (1985), que propõe uma linguística que "possua o potencial de ser aplicável em diferentes contextos e para propósitos diversos".[26] Por isso, "uma teoria se manifesta ou é 'realizada' a partir do próprio processo de ser aplicada",[27] assim, a "linguística com potencial de aplicação*"* seria "uma proposta de união entre a Linguística Teórica e a Linguística Aplicada".[28] Isso se justifica pelo fato de que uma teoria linguística sempre opera com base em opções dentro do sistema de significados de uma língua (KNIGHT; MAHBOOK, 2010, p. 5).[29] Além disso, a linguística com potencial de aplicação está "voltada ao social" (*socially accountable linguistics*), uma vez que se preocupa com questões tais quais a variação funcional (o registro) na língua e a topologia linguística, assim como uma "postura crítica" (*critical stance*) em relação a seu objeto de estudo, ou seja, em relação à língua (MATTHIESSEN, 2012).

Uma vez que esta dissertação tem como objetivo o desenvolvimento e a implementação de um conjunto de ferramentas de preparação e análise de dados estruturados (planilhas de dados) e não estruturados (do tipo texto), pode-se afirmar que esta se insere na linguística com potencial de aplicação e utiliza subsídios de outras áreas do conhecimento. A Figura 2 a seguir

[25] Minha tradução de: "[Appliable linguistics] is a kind of linguistics where theory is designed to have the potential to be applied to solve problems that arise in communities around the world, involving both reflection and action".
[26] Minha tradução de: "has the potential of being appliable in different contexts and for diverse purposes".
[27] Minha tradução de: "[…] a theory is made manifest, or 'realized', in the processes of being applied".
[28] Minha tradução de: "[...] proposed as a unifying concept to bring together theoretical and applied linguistics".
[29] "the application of a linguistic theory always operates with, or more specifically makes choices within, systems of meaning."

ilustra a relação entre a "linguística com potencial de aplicação" e as outras áreas com as quais este campo de estudo se associa nesta pesquisa.

Figura 2 – Relação entre a Linguística com potencial de aplicação e outros campos do conhecimento



Fonte: Elaborada para fins deste estudo.

Na Figura 2, pode ser observada a relação entre a Linguística com potencial de aplicação e a Ciência da Computação, a Linguística Aplicada e a Estatística, através de teorias linguísticas tais como a Linguística Sistêmico-Funcional e de subcampos da Computação – a Mineração de dados e a Mineração de textos – e da Estatística (Estatística Descritiva), incluindo técnicas de pesquisa, como as Técnicas multivariadas, descritas na subseção denominada "Estatística aplicada aos estudos linguísticos". Essas teorias, subcampos e conjuntos de técnicas são detalhados nas próximas subseções.

1.3.1 Relação da Linguística com potencial de aplicação com a Linguística Sistêmico-Funcional e a Ciência da Computação

Considerando que a Linguística com potencial de aplicação tem como pressuposto ser uma metodologia de pesquisa que deve ser utilizada de forma complementar com uma teoria linguística (KNIGHT; MAHBOOK, 2010, p. 5), é necessário utilizar como arcabouço teórico uma teoria que se mostre aplicável às situações humanas de uso, o que equivale a dizer que esta é uma teoria voltada ao social (MATTHIESSEN, 2012). Dessa forma, a Linguística Sistêmico-

Funcional pode ser considerada uma teoria linguística com potencial de aplicação, visto que define a língua como "um recurso para produzir significado [com base] [...] em padrões sistêmicos de escolhas"[30] (HALLIDAY; MATTHIESSEN, 2014, p. 23) e que se organiza em um conjunto de redes de sistemas, que podem ou não estar associados. Outra razão é o fato de que o desenvolvimento da teoria sistêmica está relacionado com sua capacidade de lidar com problemas não linguísticos, devido às diferentes correntes disciplinares que contribuíram para esta teoria e se tornaram parte dela (MATTHIESSEN, 2012, p. 438),[31] além de que "um fator de destaque na evolução da teoria sistêmica é sua permeabilidade, pois a teoria sistêmica nunca se limitou a fronteiras disciplinares" (HALLIDAY, 1985, p. 6).[32] Desse modo, por meio da Linguística com potencial de aplicação e de uma teoria "aplicável", como a Linguística Sistêmico-Funcional, é possível, primeiramente, realizar uma descrição da língua para, então, partir para sua modelagem (HALLIDAY; MCINTOSH; STREVENS, 1964). Em outras palavras, por meio da Linguística Sistêmico-Funcional, é possível desenvolver um modelo da linguagem que seja abrangente e explicativo de modo a ser aplicável tanto a problemas de pesquisa quanto a problemas práticos (HALLIDAY, 2002, p. 12).[33] Esse modelo de linguagem, como teoria, consiste, então, em um construto semiótico, sem uma disjunção entre sua teoria e sua aplicação (HALLIDAY, 2002, p. 402).[34]

Com a adoção de uma teoria abrangente da linguagem, como é a Linguística Sistêmico-Funcional, pode-se interpretar as frequências das categorias linguísticas extraídas de forma a observarmos características dos textos analisados e examinarmos escolhas linguísticas que, seja pela reiteração (reforço ou repetição de determinado tópico), seja pela ocorrência única de um determinado item, apontam para significados importantes construídos a partir da linguagem.

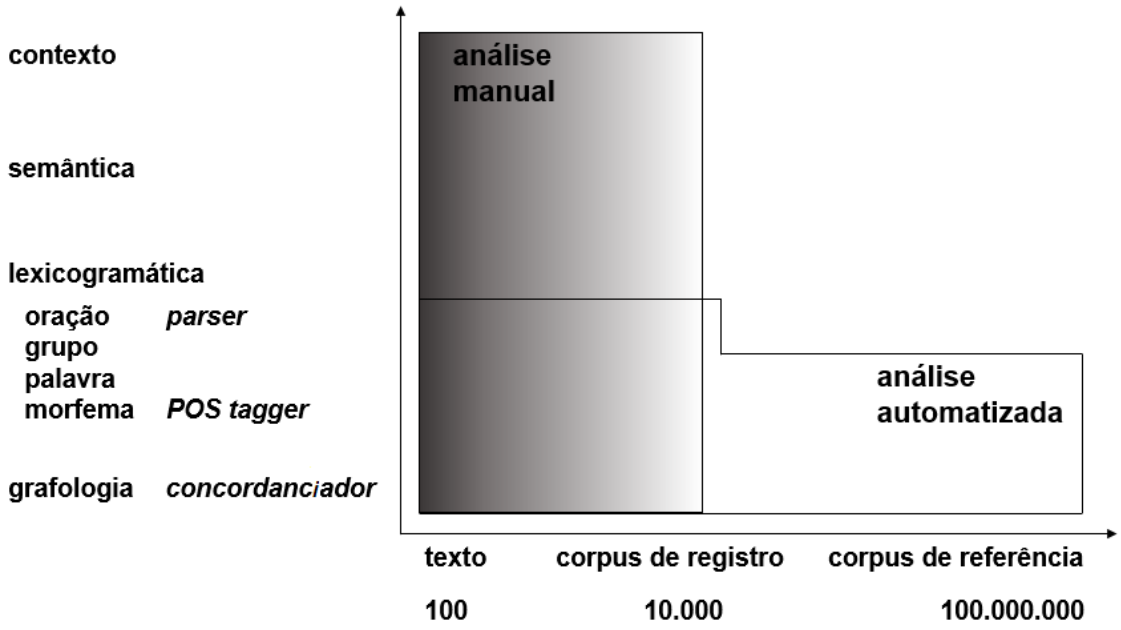
A título de exemplo, pode-se citar Castro e Cecílio (2015), que investigaram 28 entrevistas em práticas educativas com usuários do sistema de saúde com a condição crônica de diabetes, visando analisar o uso dos pronomes "eu", "nós" e "a gente" como indicadores de

[30] "A language is a resource for making meaning, and meaning resides in systemic patterns of choice."
[31] "This ability to engage with problems that lie outside linguistics itself is in fact related to the different disciplinary currents that have informed and become part of SFL, including anthropology, anthropological linguistics, sociology, educational theory, neuroscience, computational linguistics, and AI."
[32] Minha tradução de: "a salient feature in the evolution of systemic theory: its permeability from outside […] systemic theory has never been walled in by disciplinary boundaries".
[33] "[…] to bear on solving problems of a research or practical nature".
[34] "It is, as theory, a semiotic construct; but this does not create any disjunction between it and what it is theorizing – it remains permeable at all points on its surface".

responsabilidade e autonomia sobre o cuidar de si. Por meio do uso de listas de frequência, nuvens de palavras, linhas de concordância e listas de colocados, os autores observaram uma alta ocorrência do pronome “eu” (o termo mais frequente), com menor frequência dos pronomes “nós” e “a gente”. Utilizando-se a Linguística Sistêmico-Funcional para interpretar essas frequências, observou-se que, com base no sistema de SUJEITABILIDADE do sistema linguístico do português (FIGUEREDO, 2011, p. 214), formas como “nós” e “a gente” são impessoais, diferentemente de “eu”. Isso sugere que a escolha de “eu” indicaria a percepção do usuário de sua autonomia e responsabilidade sobre o cuidar de si (CASTRO; CECÍLIO, 2015).

É importante ressaltar que, embora a possibilidade de se utilizar procedimentos computacionais na análise de dados linguísticos estivesse prevista desde os anos 1950 na Linguística Sistêmico-Funcional (CATFORD, 1964; HALLIDAY, 1956, 1959, 1964) e os anos 1970 nos Estudos da Tradução (HOLMES, 1972/1988; TOURY, 1980; etc.), as condições necessárias para que procedimentos computacionais nos métodos quantitativos aplicados aos Estudos da Tradução pudessem começar a ser realizados somente foram alcançadas com a evolução da tecnologia nas últimas décadas. Assim, foi possível a utilização do computador para fazer análises em uma escala além da que um humano poderia fazer sem auxílio de ferramentas específicas. Esse potencial de análise computacional é representado na Figura 3 a seguir, que mostra um diagrama comparando o potencial da análise manual e da automatizada.

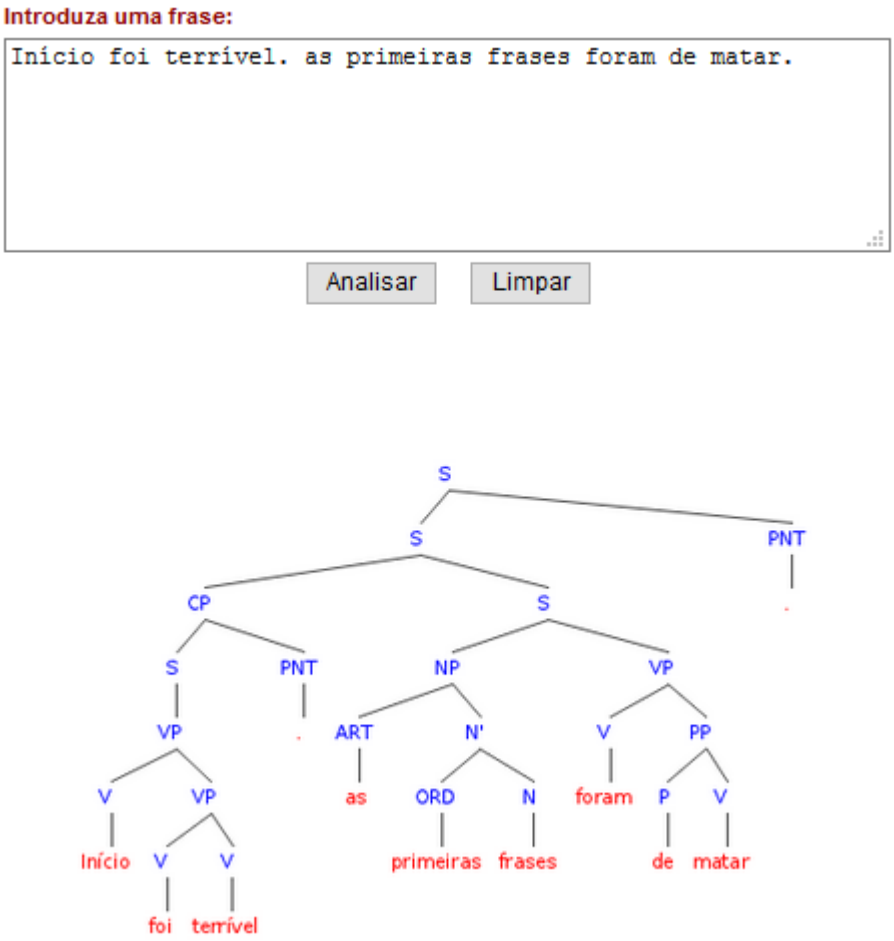
Figura 3 – Diagrama comparando o potencial da análise manual e da automatizada



Fonte: Traduzido e adaptado de Halliday e Webster (2009, p. 53).

Na Figura 3, é possível observar que no eixo vertical há categorias derivadas da Linguística Sistêmico-Funcional (grafologia, lexicogramática, semântica e contexto). Além disso, são citadas três ferramentas computacionais: o analisador sintático (*parser*), cujo exemplo pode ser visto na Figura 4, o etiquetador morfossintático (*POS tagger*) – cujo exemplo pode ser visto na Figura 5, e o concordanciador, ou melhor, a elaboração de linhas de concordância, abordada em mais detalhes no capítulo 3.

Figura 4 – Resultado gerado por um analisador sintático *on-line* de acesso livre



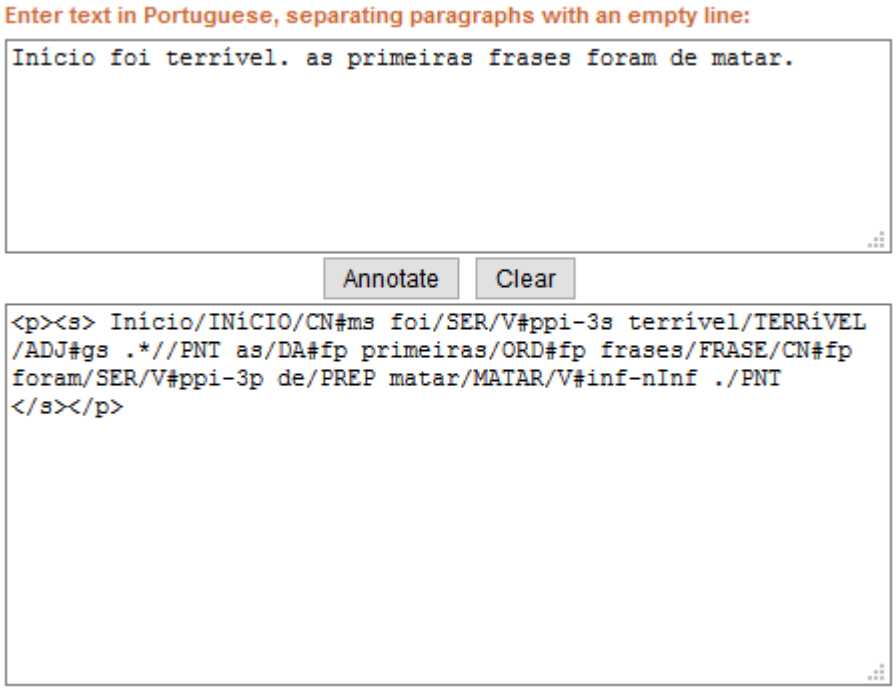
Legenda: Sentença (S), Sintagma Nominal (NP), Sintagma Verbal (VP), Sintagma Adverbial (AP), Sintagma Preposicional (PP), Artigo (ART), Substantivo (N), Verbo (V), Advérbio (ADV), Preposição (P), Sintagma Complementador (CP), Ordinais (ORD) e Pontuação (PNT). (N) é utilizado somente quando o sintagma nominal é composto de mais de um elemento.

Fonte: Elaborada com base em BRANCO *et al.* (2010).

Na Figura 4, o analisador sintático é um *software* que divide uma *string* (um texto, para a computação) em partes de acordo com categorias de gramática formal (sintagma nominais, verbais, etc.). Na oração utilizada como exemplo (“Início foi terrível. as primeiras frases foram de matar. (*sic*)”), retirada dos protocolos verbais analisados no capítulo 3, pode-se observar que as palavras foram separadas em classe de palavra por um critério morfossintático, utilizando categorias como artigo (ART), advérbio (ADV), substantivo (*noun* – N) e verbo (V). Elas foram agrupadas em sintagmas ou grupos por um critério sintático.

A seguir, na Figura 5, utilizando a mesma oração da Figura 4, segue um exemplo de uso do etiquetador morfossintático (*POS tagger*), que pode ser definido como um *software* que classifica e/ou separa as palavras de um texto pela classe de palavra (substantivo, adjetivo etc.), como no exemplo da Figura 4.

Figura 5 – Resultado gerado por um etiquetador morfossintático *on-line* de acesso livre



Fonte: Branco e Silva (2004)

Como visto na Figura 5, o etiquetador morfossintático é um *software* que, após dividir uma *string* em palavras em classes, de acordo um critério morfológico. Vale ressaltar que a anotação das categorias por parte do analisador sintático segue o seguinte formato: categoria morfossintática seguida do símbolo "#" e outras categorias. Por exemplo, a classificação de "primeiras" é: ORD#fp – ORD se refere a numerais ordinais (*ordinals*), "f" se refere a "feminino" (*feminine*) e S se refere a plural (*plural*).

Essas ferramentas, o analisador sintático e o etiquetador morfossintático permitem uma análise mais automatizada de textos até o estrato da lexicogramática, sem se estender à semântica e ao contexto. Trabalhos que utilizaram o etiquetador morfossintático em sua metodologia e que também se caracterizam como quantitativos foram Lima (2013) Ferreguetti

(2014) e Nunes (2014). Esses trabalhos contribuíram para o LETRA no sentido de apontar um movimento no laboratório em direção ao desenvolvimento de metodologias de análises estatisticamente consubstanciadas, assim como esta dissertação fornece este aporte.

O eixo horizontal da Figura 3, por sua vez, apresenta *corpora* de diferentes dimensões (do texto, com pelo menos 100 palavras, a um *corpus* de referência, com milhões de palavras), que se relacionam ao tipo de análise possível (manual ou automatizada) e com as ferramentas computacionais do eixo vertical.

A Figura 3 mostra que até determinado ponto – em que o objeto de estudo é um texto ou um *corpus* de determinado tipo de texto – é possível realizar a análise de modo totalmente manual (examinando-se tanto as categorias gramaticais quanto as semânticas). No entanto, quando são utilizados *corpora* acima de determinado tamanho, torna-se inviável fazê-lo de modo manual, sendo necessário o uso de ferramentas computacionais – utilizando-se a análise automatizada – feita com auxílio do (e não exclusivamente pelo) computador – a qual só é possível para as análises de categorias (em geral, gramaticais) passíveis de detecção pela máquina com base em parâmetros fornecidos pelo ser humano. Isso é feito pelas ferramentas citadas no eixo vertical (o *parser*, o *POS tagger* e o concordanciador).

Pode-se concluir, por meio da Figura 3, que o surgimento de novas abordagens computacionais, como a apresentada nesta dissertação, oferece benefícios para as pesquisas que fazem uso de métodos quantitativos, possibilitando que cada vez mais grande parte do trabalho antes realizado manualmente seja feito pela máquina, o que permite que o pesquisador se dedique mais a tarefas relativas a seu conhecimento de domínio, como, por exemplo, os Estudos da Tradução, assim como em Ji (2012) e Rybicky (2012). Dessa forma, com o uso de ferramentas computacionais, pode-se ampliar o potencial de análise da Linguística "aplicando" esses conhecimentos à resolução de problemas diversos; ou, em outras palavras, fazendo uso da abordagem denominada "Linguística com potencial de aplicação", a qual pode se beneficiar de outras metodologias de análise, como a Mineração de dados e a Mineração de textos, apresentadas na próxima seção.

1.3.1 Mineração de dados e Mineração de textos

Dentro da área de conhecimento da Ciência da Computação, são utilizados conjuntos de técnicas para lidar com a exploração de dados, dentre elas a Mineração de dados, que consiste no uso de técnicas computacionais para se extrair e/ou manipular grandes ou pequenos conjuntos de dados, de acordo com a definição a seguir:

> Mineração de dados (*sic*) pode ser definida como o processo de descoberta de informação aplicável a partir de grandes volumes de dados. Geralmente, essas regularidades não podem ser descobertas por meio da exploração dos dados tradicional devido ao fato de as relações serem muito complexas ou de haver um excesso de informação.[35] (SWATHI; YOGISH; SREERAJ, 2015).

Relacionada à Mineração de dados, está a Mineração de textos, a qual se diferencia da primeira no sentido de que a Mineração de dados processa dados estruturados (em tabelas numéricas organizadas em linhas e colunas, por exemplo), enquanto a Mineração de textos lida com dados não estruturados, como textos, imagens, vídeos. No caso desta dissertação, são utilizados dados não estruturados do tipo texto (*strings*).[36]

A Mineração de textos consiste em

> um vasto campo de abordagens e métodos teóricos com um ponto em comum: o texto como insumo. Isso permite diversas definições, que abrangem desde uma extensão da mineração de dados clássica voltada aos textos a formulações mais sofisticadas, como "o uso de grandes coleções de texto on-line para descobrir novos fatos e tendências sobre o próprio mundo [...]".[37] (FEINERER; HORNIK; MEYER, 2008, p. 1).

Dados estruturados podem ser definidos como dados que "organizam suas instâncias em regras bem definidas, de forma a possibilitar, através da aplicação de filtros e consultas, agrupamento e extração de dados relevantes para os usuários" (ALMEIDA, 2012, p. 8).

Dados não estruturados podem ser definidos como dados que, para serem processados, necessitam de estruturação. Isto é, "[realizar] algum procedimento que transforme a sequência

---

[35] Minha tradução de: "Data mining can be defined as the process of discovering actionable information from huge volumes of data. […]. Typically, these patterns cannot be derived by traditional data exploration because the relationships are either too complex or there prevails too much data to be dealt with."

[36] Do ponto de vista da linguística, no entanto, todo texto pode ser considerado dado estruturado, pois este consiste no próprio objeto de estudo desta disciplina, o qual será analisado tendo em vista que os itens (as palavras, os grupos, as orações, por exemplo) pertencem a categorias prédeterminadas por alguma teoria linguística, como a Linguística Sistêmico-Funcional.

[37] Minha tradução de: "a vast field of theoretical approaches and methods with one thing in common: text as input information. This allows various definitions, ranging from an extension of classical data mining to texts to more sophisticated formulations like 'the use of large on-line text collections to discover new facts and trends about the world itself […]' ".

de caracteres em objetos relacionados entre si. A lógica dessa transformação está presente no próprio texto, através de padrões linguísticos.” (ARANHA; PASSOS, 2006, p. 2). Essa visão está associada à análise computacional de textos a partir da visão da linguística, porém toda análise computacional se orienta por esses tipos de padrões.

Tomando por base a relação da Linguística de *Corpus* com os Estudos da Tradução, juntamente com a adoção de métodos quantitativos, torna-se possível a criação de uma metodologia para permitir uma análise automatizada tanto de dados estruturados quanto não estruturados.

Nas próximas duas subseções são ilustrados os conceitos de Mineração de dados e Mineração de textos apresentados acima por meio de um exemplo de dados estruturados e de um exemplo de dados não estruturados (do tipo texto).

1.3.2 Exemplo de dados estruturados

O Quadro 1 apresenta um exemplo de dados estruturados, utilizando uma amostra da planilha de dados dos pesquisadores (R1 a R4) utilizados na análise desta dissertação.

Quadro 1 – Amostra do questionário de dados preenchido pelos pesquisadores antes de realizar uma tarefa de tradução

Nome	**Proficiência em inglês [compreensão escrita]**	**Proficiência em inglês [produção oral]**	**Proficiência em inglês [produção escrita]**	**Proficiência em inglês [compreensão oral]**	**Proficiência em espanhol [compreensão escrita]**	**Número de meses de moradia em países falantes de língua inglesa**
R1	Alta	Média	Média	Alta	Alta	0
R2	Alta	Alta	Alta	Alta	Alta	8
R3	Alta	Média	Média	Alta	Alta	0
R4	Alta	Alta	Média	Alta	Alta	7

Fonte: Elaborado para fins deste estudo.

Os dados do Quadro 1 representam uma amostra dos dados do questionário respondido pelos pesquisadores (ver ANEXO 1 para um exemplo do questionário dos pesquisadores completo e ANEXO 2 para um exemplo do questionário dos tradutores completo) antes da realização de uma tarefa de tradução. Nessa amostra, as cinco colunas após a primeira (com os

códigos dos pesquisadores) representam dados de autoavaliação dos pesquisadores sobre sua proficiência linguística em inglês e espanhol na compreensão escrita, na produção oral, na compreensão oral e na compreensão escrita, além do número de meses de moradia em países falantes de língua inglesa.

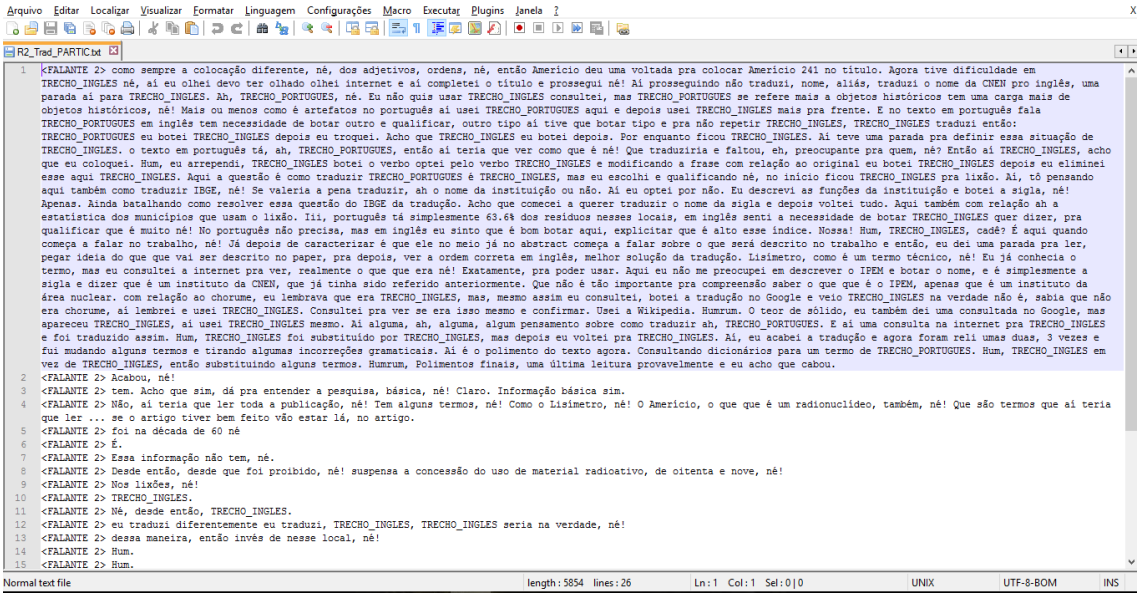
O Quadro 1, que se encontra em um formato adequado para uma análise computacional, ilustra como dados qualitativos (ou categóricos) e quantitativos podem ser estruturados em linhas e colunas. Nas linhas estão dispostos os dados das variáveis em questão, que consistem de pesquisadores participantes de experimentos – codificados com seu nome fictício (R1 a R4). As colunas representam a proficiência linguística em inglês (compreensão escrita, produção oral, produção escrita e compreensão oral), em espanhol (compreensão escrita) e número de anos de moradia em país falante de inglês. As cinco primeiras colunas possuem dados qualitativos (categóricos) e a última coluna dados quantitativos.

Retomando a diferença entre dados estruturados e não estruturados, em contraste com os dados não estruturados – que necessitam de preparação prévia específica para lidar com o objeto de estudo (no caso desta dissertação, textos), como a *tokenização* (divisão de um texto por um critério, geralmente por espaços) dos itens lexicais dos textos –, os dados estruturados geralmente necessitam de uma preparação prévia menos refinada, visto que consistem em dados numéricos. Porém, com a devida preparação, ambos os tipos de dados podem ser analisados computacionalmente.

1.3.3 Exemplo de dados não estruturados

A Figura 6 apresenta um exemplo de dados não estruturados, utilizando uma amostra de um dos textos utilizados na análise desta dissertação.

Figura 6 – Amostra de um dos textos utilizados na análise desta dissertação no *Notepad++*



A Figura 6 ilustra, na captura de tela do editor de texto *Notepad++*,[38] um exemplo de como os textos analisados nesta dissertação foram preparados para a análise, incluindo a identificação do falante (<FALANTE 2>) e os trechos citados (TRECHO_PORTUGUES e TRECHO_INGLES),[39] como será descrito em mais detalhes no capítulo da Metodologia.

1.3.4 Estatística aplicada aos estudos linguísticos

Assim como indicado na

Figura 2, pode-se utilizar a Estatística Aplicada aos estudos linguísticos como ferramenta para sumarizar e classificar os dados analisados. A Estatística Aplicada "refere-se às técnicas pelas quais os dados de natureza quantitativa são coletados, organizados, apresentados e analisados" (KAZMIER, 2004, p. 1). Dentre as técnicas estatísticas que podem ser aplicadas, encontram-se técnicas utilizadas na Estatística Descritiva, cujo enfoque são a

[38] O *Notepad++* é um editor de texto de código livre que possui suporte a diversas linguagens de programação e que funciona em MS Windows. Além disso, ele oferece suporte para a implementação de novas ferramentas por meio de *plug-ins*, assim como o uso de buscas e substituições mais avançadas utilizando, por exemplo, substituições simultâneas em um grande número de arquivos e expressões regulares (expressões linguísticas codificadas que funcionam como instruções para o computador reconhecer determinados trechos ou padrões linguísticos) (STUBBLEBINE, 2007).

[39] Nesta dissertação, devido às limpezas textuais efetuadas durante o processamento textos com base nos *scripts* no *R*, essas etiquetas (TRECHO_PORTUGUES e TRECHO_INGLES) também serão mencionadas como "trechoportugues" e "trechoingles" e como "trecho_portugues" e "trecho_ingles".

sumarização (resumo) e apresentação de dados, geralmente de uma amostra[40] da população.[41] Além disso, para análises com mais de duas variáveis (características ou categorias de interesse para o estudo) em conjunto, são utilizadas técnicas multivariadas, como a análise de conglomerados, a ser discutida em mais detalhes no capítulo seguinte.[42]

Para a sumarização dos dados utiliza-se a Estatística Descritiva, cujo objetivo é descrever de modo conciso os dados analisados de forma a caracterizar o objeto de estudo. E para classificação dos dados em grupos, utilizam-se técnicas multivariadas a fim de agrupá-los de acordo com as regularidades observadas durante a análise do tipo computacional e/ou estatística.

Quanto à aplicação da estatística na análise de um *corpus* de tradução, Oakes (2012, p. 115) afirma que há diversas maneiras de descrever um *corpus*, dentre as quais a quantificação dos recursos linguísticos. Essa quantificação pode ser feita, por exemplo, utilizando medidas de ocorrência "média" (*"average" occurrence*). Assim é possível examinar em que medida "recursos linguísticos individuais caracterizam o *corpus* como um todo" e tirar conclusões a partir da análise da variabilidade desses recursos nos textos que compõem o *corpus*.[43]

Reis (2015), que, assim como esta dissertação, apresenta os conceitos de dado estruturado e não estruturado e, portanto, utiliza conceitos da Mineração de dados e da Mineração de textos, discute as vantagens do uso de cada tipo de dado,[44] definido como "um símbolo ou um conjunto de símbolos que são usados para registrar características de um indivíduo", por meio de uma análise estatística. A autora explica que a partir dos dados é possível produzir informação, ou seja, "o produto gerado pelo tratamento de uma coleção de dados",[45] que será interpretada e gerará conhecimento,[46] que é "uma combinação de

[40] Amostra, segundo a estatística, é uma parte da população que é considerada representativa da população, ou seja, possui as mesmas características da população. Portanto, as conclusões de estudos realizados com amostras (por exemplo, de textos) podem ser generalizadas para a população dado o tratamento estatístico adequado (KERNS, 2011).
[41] População, segundo a estatística, consiste no conjunto de todas as observações sobre as quais serão retiradas amostras ou feitas inferências (KERNS, 2011).
[42] Devido ao escopo desta dissertação, só serão apresentados os "subcampos" da Estatística, segundo a Figura 2, que se aplicam a esse campo de estudo.
[43] Minha tradução de: "individual linguistic features characterise the corpus as a whole."
[44] Segundo Reis (2015, p. 38), "dado" é definido como "um símbolo ou um conjunto de símbolos que são usados para registrar características de um indivíduo".
[45] Segundo Reis (2015, p. 38), "informação" é definida como "o produto gerado pelo tratamento de uma coleção de dados".
[46] Segundo Reis (2015, p. 38), "conhecimento" é definido como "uma combinação de informação, experiência e intuição que pode beneficiar o indivíduo ou a organização (empresa)".

informação, experiência e intuição que pode beneficiar o indivíduo ou a organização (empresa)". Essa geração de conhecimento pode ser descrita pela Figura 7, a seguir:

Figura 7 – Processo para produção de conhecimento com o auxílio do tratamento estatístico dos dados

Fonte: REIS (2015, p. 39)

Como a Figura 7 mostra, o processo de produção de conhecimento utilizando a estatística para tratar os dados passa pelas seguintes etapas: (i) perguntas, (ii) estudos, (iii) dados, (iv) estatística e (v) respostas. A partir desse processo, com base nos dados, que merecem destaque em Reis (2015), é extraída informação, a qual será transformada em conhecimento.[47]

Pode-se dizer então que, assim como em Reis (2015), esta dissertação utiliza os conceitos de dado estruturado e não estruturado e faz uso desses dois tipos de dados para produzir conhecimento por meio de uma análise computacional e estatística, assim como discute as vantagens do uso de cada tipo de dado por meio de uma análise estatística.

Neste capítulo foram apresentados os principais conceitos, perspectivas e abordagens teóricas e metodológicas que nortearam o desenvolvimento da metodologia apresentada a seguir. Alguns desses conceitos foram o de Linguística com potencial de aplicação, de dados estruturados e não estruturados e de Mineração de dados e de textos, os quais contribuem para

[47] Segundo Reis (2015), citando Whitney (2007), "conhecimento" é definido como "uma combinação de informação, experiência e intuição que pode beneficiar o indivíduo ou a organização (empresa)".

a elaboração de uma metodologia quali-quantitativa, que se beneficia das vantagens das análises qualitativas e quantitativas. Dessa forma, tomando por base a relação da Linguística com potencial de aplicação e a Ciência da Computação, a Linguística Aplicada e a Estatística (ver Figura 2), juntamente com a adoção de uma abordagem quantitativa, torna-se possível a criação de uma metodologia para permitir uma análise automatizada tanto de dados estruturados quanto não estruturados, aplicável em pesquisas dos Estudos da Tradução ou de outros campos de estudo.

No próximo capítulo, é apresentada a metodologia utilizada nesta dissertação, segundo a qual são elaboradas algumas ferramentas de análise que serão exploradas mais profundamente no capítulo 3, mais especificamente a lista de frequência, a nuvem de palavras, as linhas de concordância, as listas de colocados (à esquerda e à direita do termo de busca) e o dendrograma.

CAPÍTULO 2

METODOLOGIA

Neste capítulo, são apresentados os procedimentos de preparação e os *softwares* utilizados para processamento e análise dos dados. Nas próximas seções, é apresentada a metodologia de análise de cada tipo de dado – não estruturado e estruturado – por meio da descrição da elaboração das ferramentas de análise, da aplicação dessas ferramentas e do modo como são expostos no capítulo 3 os resultados obtidos. Os dois tipos de análise são descritos em separado porque utilizam métodos diferentes, embora possuam semelhanças entre si, como mostrado em Rezende (2003).

2.1 Dados de análise

Os dados analisados nesta dissertação, tanto os estruturados quanto os não estruturados, foram obtidos em experimentos realizados no Laboratório Experimental de Tradução (LETRA), um dos principais centros de pesquisa em Estudos da Tradução. Esse centro contribui desde o ano 2000 com teses de doutorado e dissertações de mestrado para a abordagem do produto e do processo tradutórios.

Na abordagem processual, são elaborados diferentes tipos de tarefas a fim de coletar dados em condições experimentais que indiquem processamento e esforço cognitivo, "capazes de trazer à tona o que está subjacente ao ato tradutório e, assim, por meio da inferência, permitir que se acesse indiretamente o processo de tradução" (MALTA, 2015, p. 19). Para analisar esses dados, muitas vezes utiliza-se o método da triangulação de dados, com base no qual, "ao estudar dados coletados, elicitados e interpretados por meio de diferentes métodos sobre um fenômeno ou objeto de estudo, as chances de conhecer este objeto aumentam" (JAKOBSEN, 1999, p. 18).[48] Por exemplo, são dados nesse tipo de investigação os de registro de teclado e *mouse* (*key-logging*) e rastreamento ocular (*eye-tracking*) das fixações oculares dos sujeitos na tela do computador – ambos coletados por um ou mais *softwares* voltados para esse fim, como o *Tobii Studio* e o *Translog-II*.

Os experimentos dos quais os dados examinados nesta dissertação foram coletados foram realizados entre os anos de 2010 e 2012 por pesquisadores do Laboratório e tiveram como objetivo aferir o desempenho de sujeitos e depreender características da expertise

[48] Minha tradução de: "by studying data collected, elicited, and interpreted by means of different methods about the same phenomenon or object (translating), our chances of knowing this object improve".

tradutória[49] por meio de dados obtidos por diversos tipos de medição associados às tarefas realizadas. Eles foram elicitados num estudo envolvendo oito sujeitos: 4 tradutores profissionais que trabalhavam em empresas ou de forma *freelance* e 4 pesquisadores da área de engenharia nuclear que trabalham no Centro de Desenvolvimento de Tecnologia Nuclear (CDTN), sediado no *campus* Pampulha da UFMG. Alguns trabalhos do próprio LETRA que analisaram esses dados de forma parcial ou completa foram, respectivamente, Silva (em andamento) e Braga (2012).

Braga (2012) avaliou os textos traduzidos pelos tradutores profissionais e pesquisadores utilizando pareceres de uma banca de juízes, que deram notas, as quais foram utilizadas para comparar os textos traduzidos, verificando quais foram os mais bem avaliados e que critérios foram considerados mais importantes para tal seleção. Já Silva (em andamento) examina as relações de equivalência (CATFORD, 1964) identificadas entre o texto original e os 2 textos traduzidos mais bem avaliados pelos juízes em Braga (2012).

O desenho experimental da coleta cujos dados são utilizados nesta dissertação teve atividades distintas para cada grupo de sujeitos. Os tradutores realizaram: (i) um teste de cópia;[50] (ii) uma tradução direta (da L2 para a L1);[51] (iii) uma tradução inversa (da L1 para a L2); (iv) um protocolo livre, e;[52] (v) um protocolo guiado.[53] Os pesquisadores do CDTN realizaram: (i) um teste de cópia; (ii) uma tradução direta; (iii) a elaboração de um texto de popularização da ciência do resumo acadêmico; (iv) um protocolo livre, e; (v) um protocolo guiado. Todos preencheram um questionário impresso com dados sociodemográficos (como

[49] Segundo Lourenço (2007, p. 22), o conhecimento experto, em contraste com o conhecimento do tradutor novato, pressupõe "conhecimento de domínio e conhecimento discursivo" dos sujeitos ao realizarem tarefas tradutórias.
[50] O teste de cópia consiste em o sujeito copiar um texto curto não relacionado à pesquisa com o objetivo de testar características pessoais que podem interferir no experimento, como se ele/a possui alta velocidade de digitação ou se fixa o olhar no teclado com muita frequência ao digitar. Além disso, o teste de cópia também cumpre o papel de adaptar o sujeito às condições experimentais, sobretudo ao *software* utilizado, como o *Translog*, e ao leiaute do teclado.
[51] L1, L2 e L3 se referem, respectivamente, à língua materna, à segunda língua e à terceira língua de aprendizado (HAMMARBERG, 2001, p. 22), assim como informado no currículo *lattes* dos sujeitos, de onde esses dados foram retirados.
[52] Segundo Lourenço (2012, p. 62), o protocolo livre corresponde "a um relato retrospectivo realizado imediatamente após a tarefa de tradução. Nesse protocolo, orientado para a identificação do nível de metarreflexão do sujeito e fomentado, no caso de Lourenço (2012), pela reprodução cinco vezes mais rápida da tarefa tradutória na tela do Translog Supervisor©, solicitou-se aos sujeitos que verbalizassem tudo o que lhes viesse à mente em relação à execução da tarefa, como facilidades / dificuldades, estratégias e ponderações. "
[53] Segundo Lourenço (2012, p. 62), um protocolo guiado é "suscitado pela visualização de trechos em destaque no próprio texto de partida, [e] consistiu em [...] perguntas diretas que buscaram analisar o grau de entendimento dos sujeitos em relação a dois elementos coesivos (*i.e.*, máscaras, ou seja, perguntas não relacionadas diretamente ao foco do experimento) e às cinco manipulações mais metafóricas de suas respectivas versões".

formação acadêmica, experiência profissional com tradução e conhecimento linguístico) e também informações relacionadas a hábitos de leitura, escrita e tradução. Nesta dissertação, foram utilizados dados não estruturados, do tipo texto, e dados estruturados, na forma de planilhas eletrônicas.

Os dados não estruturados foram extraídos das transcrições dos protocolos verbais retrospectivos (livres e guiados) dos 8 sujeitos (os tradutores e pesquisadores do CDTN) sobre a tarefa de tradução que haviam realizado. Esses protocolos, que foram transcritos por uma pesquisadora do LETRA utilizando como critério a maior proximidade possível com o diálogo dos falantes envolvidos nas entrevistas, consistem em relatos dos sujeitos nos quais relembram suas decisões durante a tarefa realizada, assim como suas dificuldades e suas impressões gerais. Os protocolos utilizados foram os protocolos livres, nos quais o sujeito fez seu relato sem a intervenção do responsável pela coleta de dados, e os guiados, nos quais, ao contrário do que ocorre nos protocolos livres, o responsável pela coleta interveio com perguntas a fim de extrair mais informações relevantes acerca do processo de tradução. A escolha desses dados obedeceu ao objetivo pretendido: identificar padrões nos protocolos verbais dos tradutores e pesquisadores que pudessem elucidar aspectos de sua metarreflexão[54] sobre as tarefas tradutórias, objeto de investigação das pesquisas do LETRA.

Os dados estruturados utilizados nesta dissertação são os dados sociodemográficos dos sujeitos e demais dados colhidos no questionário preenchido pelos sujeitos, como hábitos de leitura e de conhecimento linguístico. Esses dados foram coletados em questionário impresso e posteriormente digitalizados e transformados em um formato eletrônico com o auxílio de uma plataforma desenvolvida para esse fim, o *Google Forms.*[55]

2.2 Escolha do *software* de análise

O principal *software* utilizado neste estudo foi o *R.* O *R,* que também é considerado um ambiente computacional, é um *software* porque consiste, em primeiro lugar, em uma linguagem de programação na qual são descritas as instruções para o computador realizar determinada tarefa. Além disso, trata-se também de um ambiente interativo onde se pode programar,

[54] Segundo Alves (2005, p. 122), metarreflexão se caracteriza pelo automonitoramento das atividades processuais durante uma tarefa de tradução.
[55] Disponível em: <https://www.google.com/forms/about/>.

utilizando recursos gráficos, ferramentas para elaborar e corrigir os *scripts*, funções nativas de pacotes[56] pré-carregados, bem como rodar programas a partir de *scripts* já prontos.[57]

A razão da escolha desse *software* é o fato de ele – ao contrário de outros *softwares* (livres ou proprietários), que também podem ser utilizados para gerar dados a partir de procedimentos usuais da Linguística de *Corpus* (como listas de frequência, listas de colocados e linhas de concordância) –, ser capaz de auxiliar não apenas na análise dos dados, mas em todas as etapas da pesquisa. Em outras palavras, o *R* auxilia nas etapas relativas aos dados (neste caso, textuais), como pré-processamento (preparação, importação, limpeza e processamento geral prévio dos dados), associação (a busca pelas associações de determinado termo a partir das contagens de frequência de co-ocorrência), aglomerados (agrupamento de itens similares em grupos) e sumarização dos dados (resumo dos padrões dos dados) (CASTRO; CECÍLIO, 2015, p. 88).

Além disso, o *R* é um *software* livre e utiliza linhas de comandos criadas ou importadas pelo usuário de outra fonte, como comunidades *on-line*, por exemplo o *Github*,[58] *R-bloggers*,[59] *stackoverflow*[60] nas quais programadores contribuem para o aprendizado de outros programadores tirando dúvidas ou até mesmo oferecendo alguns *scripts* prontos para resolver problemas enfrentados pelos usuários dessas comunidades. Em outras palavras, esse *software*, cuja interface é visualizada na Figura 8 a seguir, ao contrário de pacotes estatísticos proprietários como o *SPSS* e o *WordStat* (FEINERER; HORNIK; MEYER, 2008, p. 3), permite a livre manipulação dos dados e a customização da análise de acordo com o interesse do pesquisador, que pode utilizar quaisquer ferramentas computacionais à disposição (desde que inclusas na programação do *script*) para fazer suas análises.

---

[56] Pacotes são conjuntos de funções que podem ser importadas para uso imediato. Disponível em: <http://www.statmethods.net/interface/packages.html>. Acesso em: 24 maio 2016.

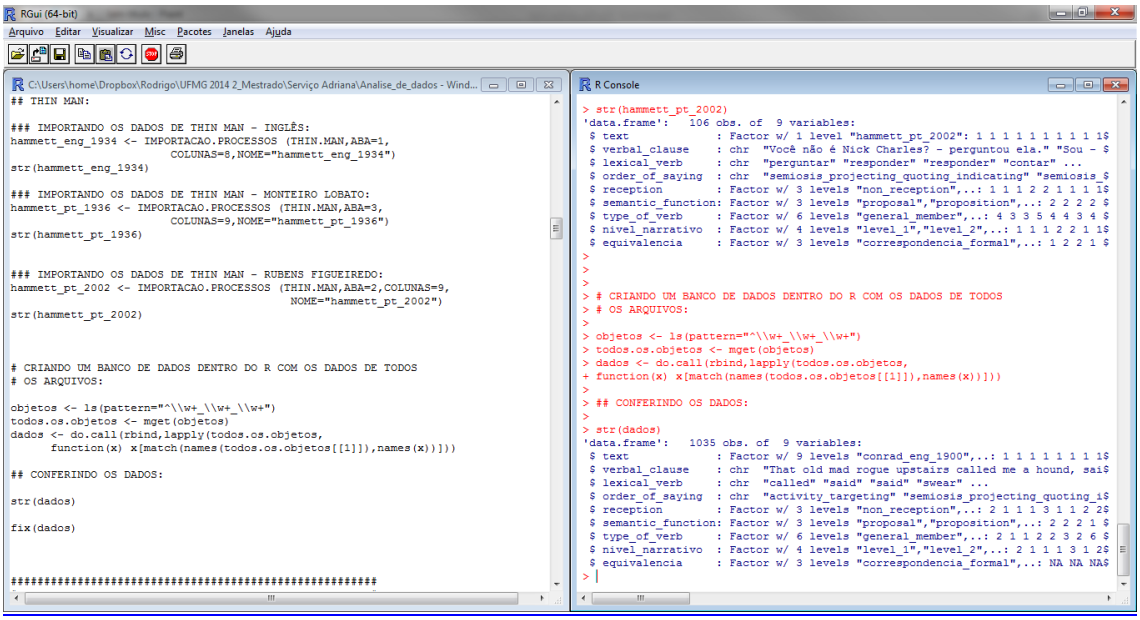
[57] "it consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs in script files." Disponível em: <https://cran.r-project.org/doc/FAQ/R-FAQ.html>. Acesso em: 13 jan. 2016.

[58] Disponível em: <https://github.com/>.

[59] Disponível em: <http://www.r-bloggers.com/>.

[60] Disponível em: <stackoverflow.com>.

Figura 8 – Captura de tela do *software* e ambiente computacional



A Figura 8 mostra uma tela capturada durante o uso do *software*/ambiente *R*, com o *script* à esquerda e, à direita, o *console* – onde o código é executado e os resultados são gerados.

Para a análise de dados desta dissertação, foram elaborados três *scripts*,[61] que consistem em arquivos de texto no formato *.R* com funções e comandos na linguagem de programação *R*. O primeiro *script* contém todas as funções elaboradas para a análise dos dados estruturados e não estruturados, assim como algumas funções auxiliares, para fazer a limpeza dos textos (eliminação de números, espaços extras etc.) ou a eliminação dos parênteses angulares (< e >). O segundo consiste no *script* de análise dos dados não estruturados, o qual, utilizando a função *source* do *R*, importa em segundo plano as funções do *script* descrito previamente. E o terceiro consiste no *script* de análise dos dados estruturados, que também importará o *script* com as funções.[62]

[61] Nos *scripts*, ao contrário desta dissertação, é necessário utilizar as aspas retas ("") a fim de evitar problemas de codificação com o *R*, *software* utilizado na análise.

[62] Os *scripts* estão disponíveis na comunidade on-line de programadores *Github*, no *link* <https://github.com/rodrigoacastro/castro2016>, especificamente nos diretórios (*branches*): <https://github.com/rodrigoacastro/castro2016/tree/functions>;<https://github.com/rodrigoacastro/castro2016/tree/analysis>. O *script* das funções se encontra no primeiro *link* e os *scripts* de análise no segundo *link*, os quais serão atualizados de tempos em tempos, de acordo com as alterações no funcionamento do *R*.

2.3 Procedimentos

2.3.1 Procedimentos de Preparação dos dados

Para que os dados pudessem ser processados pela máquina, foi necessário, em primeiro lugar, preparar semiautomaticamente os textos e as planilhas para as análises, utilizando *softwares* como editores de texto e de planilhas.

2.3.1.1 Preparação dos dados não estruturados

A preparação dos dados não estruturados consistiu, primeiramente, na conversão dos arquivos de texto do formato .doc para .txt por meio do *software* livre *Notepad++*, uma vez que a maior parte dos *softwares* que lidam com textos foi projetado para lidar com tal formato, mas não com arquivos do tipo .doc ou do tipo .pdf, assim como citado em Ferreguetti e Rodrigues (2015, p. 70). Os textos foram codificados[63] em UTF-8 para evitar, por exemplo, erros de leitura de acentos, que podem prejudicar a análise no ambiente *R*.[64]

Em seguida, utilizando recursos o *Notepad++* como o localizar e substituir e expressões regulares, foram eliminados dos protocolos analisados os trechos que representavam a fala do pesquisador responsável pela coleta dos dados, o qual interagiu com os sujeitos durante as coletas. Uma vez que se buscou mapear as falas dos sujeitos, considerou-se que as falas do pesquisador não eram de interesse para esta dissertação e que poderiam interferir nos resultados obtidos por meio das ferramentas utilizadas.

Outros procedimentos de preparação dos dados realizados no *Notepad++* consistiram na limpeza e substituição dos itens dos textos que não eram de interesse de análise ou poderiam enviesá-la,[65] seja por erros de leitura do *software* ou pela inserção de dados que, ao serem

[63] Uma codificação é uma forma utilizada pela máquina para entender os caracteres do teclado. Por isso, um arquivo de texto estará necessariamente em alguma codificação, como UTF-8 (formato padrão utilizado pelo *R*) ou ANSI (formato padrão utilizado no Windows).

[64] Outros *softwares* de análise trabalham, idealmente, com outras codificações. Por exemplo, o *AntConc*, um *software* livre, lê, por padrão, a codificação ANSI e apresenta problemas similares ao *R* se o arquivo de entrada não estiver em uma codificação adequada.

[65] Para a estatística, quando a metodologia de uma análise está enviesada, os resultados obtidos não são aqueles que melhor representam o fenômeno em estudo, o que pode levar a conclusões equivocadas. Por isso é necessário evitar utilizar dados ou métodos com potencial de enviesar a análise, visando à melhor explicação possível do objeto de estudo (LANE, 2016).

processados pela máquina, dificultam a interpretação dos resultados. Como exemplo, citam-se as aspas duplas curvas, que não são lidas adequadamente por diversos *softwares*, e os números, que influenciariam na contagem de itens e interferem em resultados gerados – como a lista de frequência.

Por isso, foram realizados, utilizando o *software Notepad++* e, quando explicitado, o *R*, os procedimentos para limpeza do texto. Algumas vezes, o *Notepad++* foi utilizado no lugar do *R* devido aos diversos recursos avançados que ele oferece, como o uso de expressões regulares, o que reduz a necessidade de elaborar *scripts* para a preparação de dados caso esses recursos já estejam disponíveis nesse *software*. Os seguintes procedimentos para limpeza do texto foram divididos em "obrigatórios" (visando, principalmente, à leitura dos dados de forma adequada) e "opcionais" (cuja motivação era permitir a análise do objeto de estudo ou evitar que esta apresentasse resultados enviesados):

Obrigatórios

- Eliminação das linhas em branco no início e no fim de cada texto, evitando, assim, que elas sejam lidas pelo *software R*;
- Edição do texto para garantir que o diálogo de cada falante esteja em apenas 1 linha, tornando mais eficiente a leitura do *software R* e facilitando o processamento posterior desses dados, ao eliminar, por exemplo, a necessidade de remover as linhas em branco em uma próxima etapa;
- Substituição de aspas duplas curvas (" e ") utilizadas no editor de texto *Microsoft Word* pelas utilizadas em editores como o *Bloco de notas* ou o *Notepad++* (");
- Eliminação de espaços extras entre palavras, que iriam interferir na segmentação dessas durante o processamento dos dados.

Opcionais

- Eliminação do cabeçalho dos textos (por exemplo, "Transcrição R2Trad:"), pois este não consiste em um dado de interesse;

- Marcação dos trechos indicadores de fala dos falantes (sujeitos e pesquisadores) entre parênteses angulares (< e >) para posterior eliminação via *script* do *R*. Exemplo: <FALANTE 1>; <FALANTE 2>;
- Eliminação de marcações da transcrição (como ININTELIGIVEL, PAUSA, RISO, RISOS), que não eram de interesse para este estudo e poderiam influenciar os resultados obtidos, por exemplo, no número de palavras total dos textos;
- Substituição dos trechos que poderiam interferir na leitura dos textos e na interpretação posterior, como as interjeições. Por exemplo, substituiu-se "Éh" por "Ééé","éh" por "ééé", "Eh" por "É" e "eh" por "é";
- Eliminação, utilizando o localizar e substituir, de trechos de palavras ou palavras completas repetidas (como “lisímetros” em “Ééé, citando aí, por exemplo, ééé, o lisímetros, lisímetros, né foi uma palavra que a princípio, né,”);
- Eliminação, utilizando o localizar e substituir, de palavras repetidas em hesitações (como “ééé” em “Hum. Ééé, *this paper describes the experiments*, ééé, cadê?”);
- Substituição de itens que haviam sido transcritos com grafias distintas (como "hamrum", “humram”, “Âhram”, "Ahrum", "ãnrum" "ânrum") por outro item (como “humrum”) a fim de que pudessem ser contabilizados uma única vez, evitando enviesar a lista de frequência;
- Substituição do item “num” por “não”, a fim de evitar problemas na contagem do “não”, o qual se mostra como um dos termos mais frequentes nos textos;
- Conversão dos números escritos por extenso (por exemplo “oitenta”) em numerais (nesse exemplo, “80”) para permitir a eliminação dos números, a fim de evitar que as ocorrências desses não aumentem nem o número de itens distintos do texto nem o número total de itens nos textos;
- Localização de trechos de textos traduzidos produzidos pelos sujeitos na tarefa de tradução e citados por eles no relato, sendo substituídos na língua inglesa por TRECHO_INGLES e em português por TRECHO_PORTUGUES a fim de contabilizá-los na lista de frequência e também evitar que esses pudessem enviesar os resultados obtidos pela inclusão de novos itens linguísticos.

2.3.1.2 Preparação dos dados estruturados

Como dados estruturados, foram utilizados os questionários dos sujeitos da pesquisa, impressos, em formato escrito, com informações sociodemográficas e outras, como hábitos de leitura e de conhecimentos linguísticos. Esses questionários foram transferidos para o meio eletrônico por meio de elaboração e preenchimento pelo autor desta dissertação no *Google Forms*, visto que as planilhas geradas seguem o formato ideal para a leitura dos dados estruturados por *softwares*. Dos arquivos eletrônicos, foram geradas duas planilhas eletrônicas com os resultados, uma com os dados dos pesquisadores e outra com os dados dos tradutores. Amostras dessas planilhas são apresentadas a seguir no Quadro 2 e no Quadro 3. As planilhas estavam em formato .csv (separadas por ponto e vírgula) e .xlsx (não utilizado na análise via *software R* para evitar problemas de leitura e processamento dos dados) e apresentavam dados relativos a informações sobre a formação prévia dos tradutores, como língua materna, anos de tradução e tipo de texto traduzido, além de dados linguísticos como níveis de conhecimento de inglês e espanhol.

Quadro 2 – Amostra da planilha de dados dos pesquisadores utilizada na análise dos dados estruturados

Nome	**Proficiência em inglês [compreensão escrita]**	**Proficiência em inglês [produção oral]**	**Proficiência em inglês [produção escrita]**	**Proficiência em inglês [compreensão oral]**	**Proficiência em espanhol [compreensão escrita]**	**...**	**Fontes de documentação [outras]**
R1	Alta	Média	Média	Alta	Alta		Não
R2	Alta	Alta	Alta	Alta	Alta		Sim
R3	Alta	Média	Média	Alta	Alta		Sim
R4	Alta	Alta	Média	Alta	Alta		Sim

Quadro 3 – Amostra da planilha de dados dos tradutores utilizadas na análise dos dados estruturados

Nome	**Graduação em engenharia**	**País do primeiro título**	**Dissertação de mestrado**	**Tese de doutorado**	**Proficiência na língua de trabalho (inglês)**	**Experiência de moradia em país falante de língua inglesa**	**…**	**Classificação das prioridades ao traduzir [resolução de problemas com base em buscas on-line e fontes impressas]**
1	Não	Brasil	Não	Não	Proficiente	Não		1
T2	Sim	Brasil	Não	Não	Bilíngue	Não		1
T3	Não	Outro	Não	Não	Proficiente	Sim		1
T4	Não	Brasil	Não	Não	Muito proficiente	Sim		1

Em seguida, os dados das planilhas foram reorganizados para importação no *R*, o *software* de análise. Essa reorganização permitiu a análise de duas formas distintas: (i) comparando os sujeitos de cada grupo separadamente; (ii) comparando os sujeitos de ambos os grupos em conjunto.

2.3.2 Metodologia de análise dos dados não estruturados

2.3.2.1 Elaboração das ferramentas de análise

Alguns dos resultados gerados por ferramentas elaboradas para a análise dos dados não estruturados na forma de *scripts* com funções do R são os seguintes:

- Lista de frequência;
- Nuvem de palavras;
- Linhas de concordância;
- Listas de colocados.

Como já citado, isso foi feito com base em pesquisas em busca de pacotes e/ou funções do R que poderiam contribuir para a elaboração dessas ferramentas. No caso dos dados não

estruturados, foram utilizados, além dos pacotes nativos do *R*, os pacotes *stringr*, *tm* e *wordcloud*, disponíveis no CRAN,[66] local onde eles ficam armazenados e a partir do qual eles podem ser acessados.

Esses pacotes são importantes porque oferecem ferramentas para complementar os recursos oferecidos pela linguagem de programação *R*, que, assim como as outras, possuem limitações. No caso do *R*, algumas limitações estão relacionadas à manipulação de textos (*strings*) e à geração de representações gráficas, como uma nuvem de palavras. Consequentemente, para lidar com *strings*, foram utilizados o pacote *stringr* e o pacote *tm* (com diversas funções relacionadas à Mineração de textos) e, para lidar com a plotagem da nuvem de palavras, foi utilizado o pacote *wordcloud*, que utiliza de funções de outros pacotes (*methods*, *RColorBrewer*, *slam* e *Rcpp*).[67]

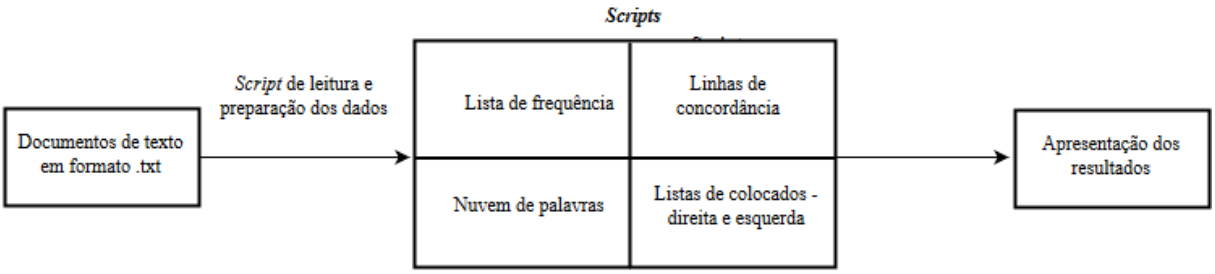
2.3.2.2 Aplicação das ferramentas

A fim de realizar a aplicação das ferramentas de análise citadas, foi necessário, primeiramente, que os dados fossem preparados adequadamente. Assim como mencionado, a preparação dos dados consistiu em procedimentos como a *tokenização* dos textos (separação das palavras em itens, neste caso utilizando o espaço como separador), a limpeza dos dados retirando o que não foi utilizado na análise, além da contagem da frequência dos itens, complementada por meio de linhas de concordância dos termos relevantes de acordo com o interesse do estudo, as quais podem ser selecionadas dentre os itens mais frequentes.

Para a análise dos dados não estruturados, na Implementação das ferramentas de análise elaboradas no capítulo seguinte, é utilizada a seguinte sequência de seguintes procedimentos, ilustradas na Figura 9

[66] Disponível em: <http://cran.r-project.org/>. Acesso em: 8 jan. 2016.
[67] Disponível em: <https://cran.r-project.org/web/packages/wordcloud/index.html>. Acesso em: 8 jan. 2016.

Figura 9 – Diagrama de análise dos dados não estruturados



Fonte: Elaborada para fins deste estudo no editor gráfico *on-line Draw.io*[68].

Como visto na Figura 9, foram utilizados os *scripts* elaborados para gerar a lista de frequência, a nuvem de palavras, as linhas de concordância e as listas de colocados para extrair resultados a partir dos dados não estruturados relevantes para esta pesquisa. Isso foi feito especificamente por meio de um *script* que importa as funções previamente elaboradas de outro *script*.

Em seguida, os resultados foram interpretados utilizando como teoria linguística a Linguística Sistêmico-Funcional, de modo a investigar indícios de metarreflexão dos sujeitos com base na análise dos PROCESSOS mentais das transcrições de seus protocolos verbais, co-ocorrendo com os PARTICIPANTES realizados ou não. Nessa análise, utilizam-se as categorias de PARTICIPANTE e PROCESSO, realizadas pelos pronomes pessoais e pelos verbos, no caso do *corpus* desse trabalho. Para possibilitar a análise dos PROCESSOS mentais, enfocaram-se, além dos pronomes pessoais "eu" e "a gente" em co-ocorrência com processos mentais, o termo "que", o qual ocorre com frequência em projeções dos PROCESSOS mentais ainda que os PARTICIPANTES não estejam realizados.

2.3.2.3 Apresentação dos resultados

A apresentação dos resultados da análise dos dados não estruturados, realizada enfocando primeiramente os dados de ambos os grupos (tradutores e pesquisadores) e em seguida cada grupo, apresentando, respectivamente, os dados dos pesquisadores e dos tradutores, foi feita de maneira numérica, na forma de tabelas, gráficos e como listas de frequência, linhas de concordância e listas de colocados. Para a elaboração e análise das tabelas

[68] Disponível em: <https://www.draw.io/>.

e listas de frequência, foram utilizados um *software* editor de planilhas e o ambiente *R*. A apresentação gráfica foi feita utilizando nuvens de palavras e dendrogramas, elaborados com base nos resultados numéricos. A nuvem de palavras foi elaborada com base em uma lista de frequência e o dendrograma com base em uma tabela numérica ou de contingência (de frequências de itens).

Para a apresentação gráfica dos textos, utilizou-se a nuvem de palavras. Reservou-se o dendrograma para a apresentação dos resultados dos dados estruturados, descritos na seção a seguir.

2.3.3 Metodologia de análise dos dados estruturados

2.3.3.1 Elaboração das ferramentas

As ferramentas utilizadas para os dados estruturados também consistem em *scripts* do ambiente *R*. Porém, devido à própria natureza dos dados analisados, não são utilizadas as mesmas ferramentas de análise dos dados não estruturados.

Devido ao fato de que o ambiente *R* se trata de um *software* estatístico, as ferramentas utilizadas consistem principalmente de *scripts* utilizando funções (como as funções *summary* e *describeBy* – essa última pertencente ao pacote *psych*) que realizam cálculos de média, mediana, moda,[69] desvio padrão,[70] intervalos de confiança[71] e tabelas sumarizando os dados encontrados para que a análise dos mesmos possa ser feita. No entanto, foram elaborados *scripts* para a análise por meio de técnicas multivariadas, com a análise de conglomerados, por meio de dendrogramas (cf. ALBUQUERQUE, 2005).

[69] Para a estatística, a média, a mediana e a moda são consideradas medidas de tendência central, ou seja, medidas para sumarizar os dados encontrando um valor médio que os representa. A média pode ser definida como a soma entre todos os valores divididos pelo número de valores. A mediana é o valor que se encontra exatamente no meio de um conjunto de valores (ordenados em ordem crescente). E a moda é o valor mais frequente dentre um conjunto de dados (KERNS, 2011).

[70] Na estatística, o desvio padrão representa uma medida da variação (desvio) em relação à média, visto que toda medida se encontra no centro de um intervalo, com valores abaixo e acima do valor central.

[71] Na estatística, intervalo de confiança é uma faixa de valores em volta de um valor central (geralmente a média ou a mediana), cuja amplitude (diferença entre valor máximo e mínimo) é determinada por uma tolerância de erro a critério do pesquisador (KERNS, 2011).

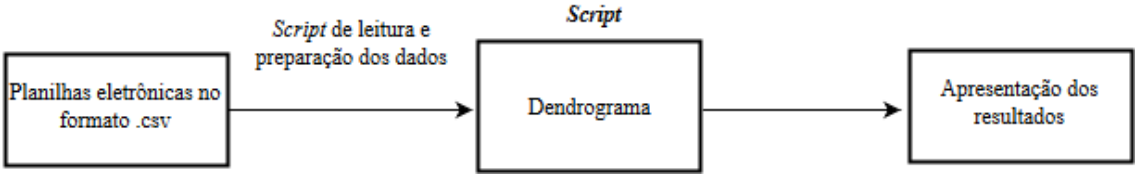
No caso dos dados estruturados, foram utilizados, além dos pacotes nativos do *R*, os pacotes *psych* e *stats*. Ambos os pacotes possuem ferramentas estatísticas, como a função *describeBy* do *psych* para sumarizar os dados.

2.3.3.2 Aplicação das ferramentas

Para a aplicação dessas ferramentas, também foi necessário, assim como no caso dos dados não estruturados, que algumas adequações fossem feitas a fim de que os dados assumissem o formato apropriado para serem explorados. No entanto, as operações realizadas não são as mesmas, pois nesse caso a reorganização dos dados é feita principalmente dentro do próprio ambiente *R*. Isso é feito por meio de um *script* que importa de outro arquivo as funções previamente elaboradas. No caso, será utilizada basicamente a função para fazer o dendrograma.

A Figura 10, a seguir, resume o processo de análise dos dados estruturados, desde as planilhas preparadas previamente e convertidas no formato .csv até a apresentação dos resultados.

Figura 10 – Diagrama de análise dos dados estruturados



Fonte: Elaborada para fins deste estudo no editor gráfico *on-line Draw.io.*

Após as etapas descritas na Figura 10, os resultados foram relacionados aqueles dos dados não estruturados, os quais foram interpretados utilizando a Linguística Sistêmico-Funcional, de modo ao analisar os pronomes pessoais “eu” e “a gente” e os PROCESSOS mentais os quais com eles co-ocorrem.

2.3.3.3 Apresentação dos resultados

Para os resultados dos dados estruturados, realizaram-se a sumarização e o agrupamento dos dados para pudessem ser apresentados de forma numérica e gráfica. Além disso, assim como já citado, o dendrograma foi utilizado para a apresentação gráfica dos resultados dos dados estruturados.

No próximo capítulo, será apresentada a implementação das ferramentas de análise elaboradas, realizada com os dados da análise por meio dos procedimentos descritos neste capítulo.

CAPÍTULO 3

IMPLEMENTAÇÃO DAS FERRAMENTAS DE ANÁLISE ELABORADAS

A implementação das ferramentas de análise elaboradas, apresentada neste capítulo, descreve a aplicação da metodologia de pesquisa de dados estruturados e não estruturados mostrada no capítulo 2. Também discute os resultados gerados, apontando as conclusões e outras possibilidades de análise integrando o uso de dados estruturados e não estruturados em uma abordagem quali-quantitativa utilizando os métodos quantitativos, isto é, uma abordagem quantitativa fazendo uso de ferramentas computacionais e estatísticas. Esses dois tipos de análise, que, como já mencionado, utilizam métodos diferentes, porém possuem semelhanças entre si.

Para dar consecução ao objetivo desta dissertação, qual seja desenvolver, implementar e testar um conjunto de ferramentas de preparação e análise de dados estruturados e não estruturados utilizando o *software* estatístico e ambiente computacional *R*, foi realizado um estudo visando aplicar as ferramentas desenvolvidas por meio dos *scripts*. Além disso, foi feita a triangulação dos resultados encontrados para obter conclusões que derivem da complementaridade dos dados qualitativos e quantitativos analisados.

Como foi exposto, no escopo de uma *Linguística com potencial de aplicação*, desenhada no marco teórico da linguística sistêmico-funcional e que, neste trabalho, utiliza subsídios da Linguística de *Corpus*, da Mineração de dados e de textos, da Estatística Descritiva e do uso de técnicas multivariadas de análise, foi possível desenvolver ferramentas de análise computacional que podem ser aplicadas a diversos campos disciplinares – por exemplo, aos Estudos da Tradução. Nesse campo, pelo ponto de vista do produto da tradução, o interesse se volta à linguagem utilizada em textos traduzidos e ao processo cognitivo dos tradutores ao realizar uma tarefa tradutória, dentre outros.

Nesta dissertação, no caso de dados não estruturados, foram analisadas as ocorrências dos pronomes pessoais “eu” e “a gente” nos protocolos dos sujeitos do experimento e os verbos que co-ocorrem com esses pronomes. Sob a perspectiva da Linguística Sistêmico-Funcional, são utilizadas para a análise as categorias de PARTICIPANTE e PROCESSO, pertencentes ao sistema de TRANSITIVIDADE na metafunção ideacional (ordem da oração) e realizadas pelos pronomes e pelos verbos, respectivamente. Essas categorias foram consideradas as mais relevantes para a análise e para o estudo de implementação das ferramentas elaboradas apresentado neste capítulo.

A fim de se obter evidências de metarreflexões e decisões tomadas pelos sujeitos durante a execução da tarefa com base em suas verbalizações (ALVES, 2003), tomou-se como pressuposto que esses sujeitos são representados na linguagem pelos PARTICIPANTES realizados pelos pronomes "eu" e "a gente" e pelos PROCESSOS, principalmente o PROCESSO mental – que apresentaria de forma "mais deliberada" a metarreflexão (MAGALHÃES; ALVES, 2006). Assim, essas verbalizações, na forma de protocolos verbais (livres e guiados), podem ser interpretadas numa perspectiva discursiva com o objetivo de se estudar o papel desses participantes utilizando também os tipos de PROCESSO com os quais estão associados.

3.1 Dados de análise

Como mencionado no capítulo da Metodologia, os dados utilizados são provenientes de experimentos realizados no LETRA com quatro pesquisadores do CDTN e quatro tradutores profissionais, os quais preencheram um formulário com dados sociodemográficos e outras informações, tais como hábitos de leitura e conhecimentos linguísticos. Posteriormente, foram analisados dados previamente registrados de protocolos verbais (livres e guiados) da tarefa realizada. Esses dados utilizados nesta dissertação são, portanto, do tipo estruturado e não estruturado, respectivamente.

Como visto na Metodologia, os dados de ambos os tipos (estruturados e não estruturados) foram preparados a fim de possibilitar a análise. A preparação dos dados foi feita por meio dos *softwares Notepad++* e do ambiente computacional *R*, por meio da elaboração de 3 tipos de *scripts*: para a documentação (e posterior importação) das funções utilizadas; para a análise de dados não estruturados; e de dados estruturados – disponíveis na comunidade on-line *Github* nos *links* já informados na Metodologia. O objetivo dos *scripts* foi, para além de realizar, quando necessária, a limpeza da pontuação e/ou dos números, realizar também o processamento dos dados para prepará-los para as ferramentas de análise utilizadas e permitir a geração de resultados – a lista de frequência, a nuvem de palavras, as linhas de concordância, as listas de colocados (para os dados não estruturados) e o dendrograma (para os dados estruturados). No caso dos dados estruturados, alguns procedimentos realizados para permitir o processamento computacional foram a digitalização dos questionários via *Google Forms* e a preparação desses questionários em formato digital no ambiente *R*.

Para os dados não estruturados, dentre os procedimentos de preparação dos dados, inseriu-se a marcação da indicação de cada falante entre parênteses angulares e dos trechos dos resumos acadêmicos citados pelos sujeitos, considerados dados não relevantes, além da seleção de apenas as falas dos sujeitos, com a eliminação das falas dos pesquisadores que conduziram os experimentos, também considerados não relevantes.

3.2 Resultados

3.2.1 Dados não estruturados

3.2.1.1 Lista de frequência

Para fazer uso da primeira ferramenta de análise dos dados não estruturados com o propósito de extrair dados quantitativos dos protocolos e analisar possíveis indicadores de metarreflexão dos sujeitos, representados na linguagem por meio dos pronomes “eu” e “a gente” e por meio dos PROCESSOS (principalmente o PROCESSO mental), os textos foram preparados para análise com procedimentos realizados no *Notepad++* e no *R*. Além disso, foi necessário fazer uma substituição textual para permitir a busca desejada (“a gente” por “a-gente”) – procedimento que não interferiu na análise. Essa substituição foi feita para evitar erros na etapa de preparação dos textos via *scripts*, que envolve a segmentação dos textos por espaço.

Posteriormente, foi elaborada uma lista de frequência de todos os termos dos textos analisados do *corpus* e, em seguida, foram eliminados aqueles que não eram de interesse por meio de uma lista de *stopwords*, que inclua todos os termos diferentes de “eu”, “não”, “a gente” e os verbos na primeira e na terceira pessoa do singular. Como resultado, foi gerada a lista de frequência apresentada na Tabela 1.

Tabela 1 – Lista de frequência dos vinte primeiros termos mais relevantes para a análise de “eu” e “a gente”

Termo	Frequência
eu	463
não	233
é	221
tem	60
foi	53
acho	38
era	27
achei	24
tava	24
tive	24
ficou	22
tô	22
fui	21
tinha	21
fiquei	20
sei	19
vou	18
traduzi	17
coloquei	15
a-gente[72]	14

Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Tabela 1, podem-se observar, em destaque, os pronomes “eu” (com frequência de 463) e “a gente” (com frequência de 14). É importante destacar que não houve ocorrência de outros pronomes pessoais com função de sujeito dentre os 20 termos relevantes mais frequentes, como o pronome “nós”, visto que esse pronome ocorreu apenas 2 vezes nos protocolos analisados.

Outro termo que pode ser destacado é “não”, que, quando ligado a um grupo verbal, altera a POLARIDADE do mesmo e da oração. Os demais itens são verbos conjugados na primeira ou na terceira pessoa do singular, possivelmente relacionados a ações dos PARTICIPANTES.

[72] Como pôde ser observado na lista de frequência durante sua limpeza, a real frequência de “a gente” foi de 19, visto que esse pronome ocorreu em outras formas. Essas outras formas foram “da gente” (com frequência 2), “pra gente” (também com frequência 2) e “para gente” (com frequência 1), que não se destacaram na lista de frequência devido a sua baixa frequência.

Nessa lista de frequência, pode-se observar também a ocorrência de verbos que realizam PROCESSOS materiais (como “traduzi” e “coloquei”), mentais (como “sei”) e relacionais (como “tem”, “era”, “ficou” e “fiquei”). A frequência de “a gente”, relacionada aos verbos conjugados na terceira pessoa do singular, mostra que em alguns casos os sujeitos (seja pesquisador ou tradutor) utilizam um pronome impessoal em vez de um pronome responsável, deixando de atribuir responsabilidade modal a si mesmos (cf. FIGUEREDO, 2011, p. 194) ao utilizar verbos que realizam PROCESSOS em suas verbalizações nos protocolos verbais (CASTRO; CECÍLIO, 2015).

Por fim, os PARTICIPANTES “eu” e “a gente” e os tipos de PROCESSOS utilizados, como “traduzi” e “coloquei”, parecem indicar os procedimentos realizados pelos sujeitos durante a tarefa de tradução.

Com base na Tabela 1, seguem duas listas de frequência relativas ao grupo dos pesquisadores e dos tradutores, apresentadas na Tabela 2 e na Tabela 3.

Tabela 2 – Lista de frequência dos vinte primeiros termos mais relevantes para a análise de “eu” e “a gente” no grupo dos pesquisadores

Termo	Frequência
eu	277
não	164
é	159
trechoingles	115
trechoportugues	52
tem	41
foi	38
acho	25
era	22
tive	21
achei	16
tô	16
sei	14
tava	13
a-gente	11
falei	11
traduzi	11
comecei	10
fiquei	10
fui	10

Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Tabela 2, pode-se observar que, além dos termos “eu”, “não” e “a-gente”, encontram-se em destaque “trechoingles” e “trechoportugues”, omitidos na Tabela 1 para permitir a apresentação de “a gente”. Esses termos, com frequências 115 e 62, representam as citações dos trechos do resumo acadêmico em inglês e em português brasileiro, os quais foram contabilizados para análise.

Tabela 3 – Lista de frequência dos vinte primeiros termos mais relevantes para a análise de “eu” e “a gente” no grupo dos tradutores

Termo	Frequência
eu	187
trechoingles	79
não	69
tem	19
ficou	18
foi	15
acho	13
fui	11
tava	11
tinha	11
coloquei	10
fiquei	10
trechoportugues	10
vou	10
achei	8
vai	7
consultei	6
deixei	6
fiz	6
tô	6

Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Tabela 3, assim como na Tabela 2, podem-se observar os termos “eu” e “não” em destaque, assim como “trechoingles” (com frequência 79) e “trechoportugues” (com frequência 10) se referindo às citações. Comparando-se à Tabela 2, esses termos referentes às citações nos protocolos verbais apresentaram frequência mais baixa que nas transcrições dos pesquisadores (nas quais as frequências foram 115 e 62).

Comparando a Tabela 2 e a Tabela 3, quanto à frequência do pronome “eu”, de “trechoingles” e “trechoportugues” (visto que “a gente” não ocorreu na Tabela 3), levou-se em conta a frequência relativa em relação ao número de *tokens* para cada grupo. Assim, foi possível fazer uma comparação entre os grupos por meio da Tabela 4.

Tabela 4 – Comparação entre frequências absolutas e relativas de "eu", "trechoingles" e "trechoportugues" entre o grupo dos pesquisadores e dos tradutores

Termos	**Pesquisadores**		**Tradutores**	
	Frequência absoluta	*Frequência relativa*	*Frequência absoluta*	*Frequência relativa*
"eu"	277	5,15	187	6,63
"trechoingles"	115	2,14	79	2,80
"trechoportugues"	52	0,97	10	0,35
Total	**444**	**8,25**	**276**	**9,79**
Total nos textos	**5383**	**100**	**2820**	**100**

Fonte: Elaborada para fins deste estudo. Destaque meu.

No caso do pronome "eu" e de "trechoingles", obteve-se uma frequência relativa (calculada pela frequência absoluta dividida pelo número de *tokens* de cada grupo) maior no grupo dos tradutores (6,63% e 2,80%) em relação ao número de *tokens* dos textos, porém a frequência relativa de "trechoportugues" foi maior nos textos dos pesquisadores (0,97%). Isso sugere que, em relação ao pronome "eu", ocorre maior nível de metarreflexão no grupo dos tradutores e também que, ao contrário dos pesquisadores, os tradutores citaram mais trechos dos insumos (textos usados nas tarefas) em inglês que em português nos protocolos verbais.

A Tabela 5 e a Tabela 6 a seguir, elaboradas a partir da Tabela 2 e da Tabela 3, apresentam os PROCESSOS mentais presentes nas transcrições do grupo dos pesquisadores e dos tradutores. Foi enfocado esse tipo de PROCESSO levando em conta o pressuposto de que esse tipo de PROCESSO sugere de forma "mais deliberada" (MAGALHÃES; ALVES, 2006) a metarreflexão dos sujeitos.

Tabela 5 – Lista de frequência dos processos mentais do grupo dos pesquisadores

Termo	**Frequência**
achar	54
quer	20
saber	18
entender	8
gostar	8
esperar	7
esquecer	7
olhar	6
optar	6
escolher	6
lembrar	5
conhecer	4
pensar	3
resolver	3
avaliar	2
considerar	2
sentir	2
preocupar	1
ver	1
arrepender	1
enxergar	1
precisar	1
raciocinar	1

Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Tabela 5 observam-se os PROCESSOS mentais dos protocolos verbais dos pesquisadores, totalizando 23 itens distintos. Desses, os mais frequentes são “achar” (54), “quer” (20) e “saber” (18) e os menos frequentes (com frequência 1), são “preocupar”, “ver”, “arrepender”, “enxergar”, “precisar” e “raciocinar”. Essa tabela pode ser comparada com a Tabela 6 a seguir, que foi elaborada a partir dos protocolos dos tradutores.

Tabela 6 – Lista de frequência dos processos mentais do grupo dos tradutores

Termo	Frequência
achar	9
preferir	7
saber	4
lembrar	3
quer	2
ver	2
acreditar	1
conhecer	1
descobrir	1
esquecer	1
gostar	1
olhar	1
optar	1
preocupar	1
resolver	1

Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Tabela 6 observam-se os PROCESSOS mentais dos protocolos verbais dos tradutores, representados por 15 itens distintos. Desses, alguns dos verbos realizando os PROCESSOS também estão presentes na Tabela 5, com base nos protocolos dos pesquisadores, como é mostrado comparativamente na Tabela 7.

Comparando a Tabela 5 e a Tabela 6, foi possível observar que há 23 tipos de PROCESSOS mentais distintos para os pesquisadores e 15 distintos para os tradutores. Isso pode ser devido tanto a fatores como o tipo de atividade profissional de cada grupo quanto à diferença no número de *tokens* dos textos de cada grupo (5383 nos textos dos pesquisadores e 2820 nos textos dos tradutores).

Dividindo o número de processos mentais distintos pelo número de *tokens* de cada grupo, é possível obter os índices de 0,0053191 (15/2820) para os textos dos tradutores e 0,0042727 (23/5383) para os pesquisadores. Isso mostra que, proporcionalmente, os tradutores utilizam mais PROCESSOS mentais que os pesquisadores, sugerindo maior grau de metarreflexão no grupo dos tradutores, embora os pesquisadores parecem apresentarem maior número de PROCESSOS mentais diferentes.

A seguir, com base no fato de alguns dos verbos realizando os PROCESSOS na Tabela 6 também estarem presentes na Tabela 5, foi possível elaborar a Tabela 7, que mostra comparativamente os PROCESSOS mentais comuns para ambos os grupos.

Tabela 7 – Comparação das listas de frequência dos processos mentais realizados nos protocolos verbais dos pesquisadores e dos tradutores

Termo	**Frequência-Pesquisadores**		**Frequência-Tradutores**	
	Absoluta	**Relativa (%)**	**Absoluta**	**Relativa (%)**
achar	**54**	**40,60**	9	33,30
quer	**20**	**15,04**	2	7,41
saber	**18**	**13,53**	4	14,81
gostar	**8**	**6,02**	1	3,71
esquecer	**7**	**5,26**	1	3,71
olhar	**6**	**4,51**	1	3,71
optar	**6**	**4,51**	1	3,71
lembrar	**5**	**3,76**	3	11,1
conhecer	**4**	**3,01**	1	3,71
resolver	**3**	**2,26**	1	3,71
ver	1	0,75	**2**	**7,41**
preocupar	1	0,75	1	3,71
Total	133	100	27	100

Fonte: Elaborada para fins deste estudo. Destaque meu.

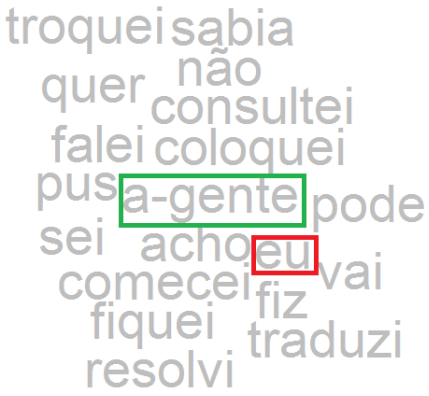
Na Tabela 7, que compara as listas de frequência dos verbos realizando processos mentais em comum utilizados pelos tradutores e pelos pesquisadores, podem-se verificar as frequências absolutas e relativas de cada processo em relação ao total de processos mentais em cada grupo.

Com base nisso, podem-se fazer comparações sobre os processos mais utilizados em cada grupo (destacados em negrito). Por exemplo, os verbos “achar”, “quer” e “saber”, assim como diversos outros, são mais frequentes no grupo dos pesquisadores que no dos tradutores, enquanto “ver” apresenta um comportamento contrário e “preocupar” é tão frequente em um grupo quanto no outro.

3.2.1.2 Nuvem de palavras

Utilizando-se como base a lista de frequência da Tabela 1, foi gerada uma nuvem de palavras, apresentada na Figura 11 abaixo.

Figura 11 – Nuvem de palavras relativa à lista de frequência dos 20 termos mais relevantes da análise

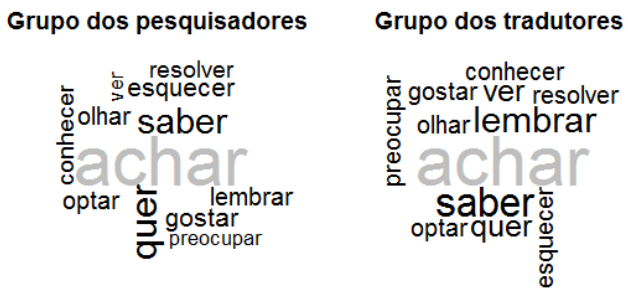


Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Figura 11, podem ser observados que os termos “eu” e “a gente”, em destaque, se encontram próximos do centro, indicando sua frequência mais alta que outros termos. Isso está de acordo com a lista de frequência da Tabela 1, considerando que a nuvem de palavras é, na verdade, uma forma de representar visualmente uma lista de frequência.

Devido à frequência desses dois termos (“eu” e “a gente”) e sua relação com a ocorrência de outros verbos, a qual pode ser interpretada por meio de uma análise das categorias de PARTICIPANTE e PROCESSO utilizando a Linguística Sistêmico-Funcional, foram os escolhidos os verbos, especificamente os realizando PROCESSOS mentais (considerados mais representativos da metarreflexão) para a análise com as ferramentas de linhas de concordância, cujos resultados são comparados com as listas de colocados de “eu” e “a gente”. A seguir, é apresentada a Figura 12, com a comparação das nuvens de palavras dos pesquisadores e dos tradutores, elaboradas respectivamente com base na Tabela 5 e na Tabela 6.

Figura 12 – Captura de tela do *R* com a nuvem de palavras relativa à lista de frequência dos vinte processos mentais distintos da análise dos pesquisadores (à esquerda) e tradutores (à direita) com frequência mínima 1

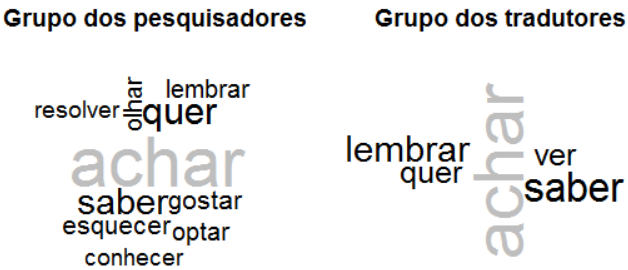


Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Figura 12, elaborada considerando-se como frequência mínima de ocorrência dos verbos realizando PROCESSOS mentais a frequência 1, é possível observar que as nuvens de palavras mostram todos os 12 PROCESSOS mentais utilizados pelos pesquisadores e pelos tradutores. Embora a distribuição dos processos seja distinta nas nuvens, os termos listados (inclusive "achar", em destaque no centro de ambas) são os mesmos, pois a lista de frequência utilizada com base foram os PROCESSOS mentais utilizados por ambos os grupos de sujeitos.

No entanto, por meio da Figura 12, não é possível contrastar os grupos com base na análise dos verbos realizando PROCESSOS mentais, visto que ambas as nuvens apresentam os dados. Por isso, visando a comparação das nuvens, segue a Figura 13, com a comparação das nuvens de palavras dos pesquisadores e dos tradutores, também elaboradas com base na Tabela 5 e na Tabela 6.

Figura 13 – Nuvem de palavras relativa à lista de frequência dos 20 processos mentais distintos da análise dos pesquisadores (à esquerda) e tradutores (à direita) com frequência mínima 2



Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Figura 13, em contraste com a Figura 12 – que utilizou como frequência mínima de ocorrência dos verbos realizando PROCESSOS mentais a frequência 1 –, utilizou-se a frequência 2. Com essa escolha, eliminou-se a ocorrência dos *hapax legomena*, ou seja, os termos cuja frequência absoluta equivale a 1, considerados de baixa frequência (SARDINHA, 2000, p. 344).

Contrastando a Figura 12 e a Figura 13, observa-se que, na Figura 12, ambas as nuvens apresentaram 12 PROCESSOS mentais e, na Figura 13, a nuvem de palavras dos pesquisadores apresentou 10 PROCESSOS e a dos tradutores 5 PROCESSOS mentais (dos 12 em comum entre os grupos). A maior variedade de PROCESSOS mentais por parte dos pesquisadores pode indicar que eles apresentaram maior metarreflexão que os tradutores, possivelmente devido ao fato de o texto utilizado no experimento ser do conhecimento de domínio dos pesquisadores. Em outras palavras, isso está relacionado ao CAMPO, uma variável no estrato do CONTEXTO localizada na metafunção ideacional, componente lógico, que representa o assunto do texto (MATTHIESSEN; TERUYA; LAM, 2010, p. 106) – no caso dos textos dos experimentos, engenharia nuclear.

3.2.1.3 Linhas de concordância

Com o objetivo de obter evidências de metarreflexão nos textos dos sujeitos do grupo dos pesquisadores e dos tradutores por meio da análise dos PROCESSOS mentais associados aos pronomes “eu” e “a gente” na função de SUJEITO, foram geradas linhas de concordância. Porém, uma vez que esses pronomes podem ou não estar realizados no português brasileiro, utilizou-se o termo de busca “que”, utilizado juntamente com PROCESSOS mentais a fim de representar a projeção (HALLIDAY; MATTHIESSEN, 2014, p. 253), como em “penso que”. Foi utilizado esse termo de busca porque caso fossem utilizadas apenas as linhas de concordância de “eu” e “a gente”, não teria sido possível contabilizar os PROCESSOS não realizados.

Assim, por meio de uma das funções do *R* elaboradas neste estudo, foram geradas as linhas de concordância dos PROCESSOS mentais de cada grupo de participantes por meio da obtenção das linhas de concordância do termo de busca “que”. Nessas linhas, o termo de busca se encontra no centro, com 5 termos à esquerda (A1 a A5) e 5 à direita (D1 a D5). Em seguida, fez-se uma filtragem semiautomática (utilizando funções do *R* e planilhas eletrônicas) selecionando apenas as linhas que apresentavam projeção. Considerou-se que essas linhas eram aqueles nas quais havia o termo “que” na coluna “Termo de busca” e um PROCESSO mental na

posição imediatamente anterior (neste caso, na coluna A5 das planilhas apresentadas nessa seção). Além disso, a fim de possibilitar a comparação entre os grupos, foram utilizados os PROCESSOS mentais comuns entre ambos os tradutores e os pesquisadores (ver Tabela 6) para a apresentação dos resultados, dos quais foi feita uma amostragem de 1 linha de cada PROCESSO. E com base nessa amostragem foram removidas no *script* as linhas que não apresentavam projeção.

Como resultado, antes da filtragem das linhas de concordância obtidas com o "que" como termo de busca, foram geradas 264 linhas para o grupo dos pesquisadores e 93 linhas para o grupo dos tradutores. A filtragem dessas linhas foi feita utilizando listas de frequência geradas no *R* das linhas que apresentavam PROCESSOS mentais conjugados na primeira ou terceira pessoa do singular e que apresentavam projeção.

Após a filtragem semiautomática das linhas de concordância do termo de busca "que" para cada grupo, foram obtidas 7 linhas de concordância indicando projeção para os tradutores e 42 linhas para os pesquisadores. Essas novas linhas apresentavam diversos PROCESSOS mentais distintos, como mostrado na Tabela 5 (com 23 verbos diferentes) dos pesquisadores e na Tabela 6 (com 15 verbos diferentes) dos tradutores. Com base nessas linhas com PROCESSOS mentais na primeira ou na terceira pessoa do singular indicando projeção, após a amostragem de 1 linha por verbo e da eliminação daquelas que não apresentavam projeção, foram elaborados os quadros a seguir.

Quadro 4 – Linhas de concordância aleatórias de "que" dos processos mentais comuns para o grupo dos pesquisadores

A1	A2	A3	A4	A5	**Termo de busca**	D1	D2	D3	D4	D5
mais	de	um	dicionário.	acho	**que**	é	só.	é	não	foi
eu	quer	dizer	eu	achava	**que**	eu	estava	batendo,	né,	normal
na	verdade	não	é,	sabia	**que**	não	era	chorume,	aí	lembrei
trecho_ingles	or	trecho_ingles	e	achei	**que**	trecho_ingles	era	melhor,	mas	eu
lixão.	mas	eu	só	lembrei	**que**	eu	podia	usar	isso	no

Fonte: Elaborado para fins deste estudo. Destaque meu.

O Quadro 4 apresenta cinco linhas de concordância dos pesquisadores do termo de busca "que" selecionadas aleatoriamente dentre as linhas com verbos realizando processos mentais com projeção. Nota-se que os processos mentais a 1 posição à esquerda (coluna A5) do termo

de busca, como “acho”, “sabia” e “lembrei”. Além disso, em uma dessas ocorrências, tem-se “não” ligado ao grupo verbal, marcando a POLARIDADE negativa dessas orações.

Analisando a ocorrência dos PROCESSOS mentais (“acho”, “achava”, “sabia”, “achei” e “lembrei”), destaca-se que todos eles apresentam projeção, sendo caracterizados como metáforas interpessoais[73] (como “acho que” e “sabia que”). Pode-se dizer que os processos mentais, assim como os que caracterizam metáforas interpessoais, indicam metarreflexão por meio dos sujeitos, o que pode ser comprovado pela recorrência de “trecho_ingles” co-ocorrendo com o PARTICIPANTE “eu” à direita e “eu” co-ocorrendo à direita e à esquerda, como observado no Quadro 4.

Quadro 5 – Linhas de concordância aleatórias de “que”
dos processos mentais comuns para o grupo dos tradutores

A1	A2	A3	A4	A5	Termo de busca	D1	D2	D3	D4	D5
de	usinas,	onde	eu	sabia	**que**	existia	por	exemplo,	para-raios,	né.
ali	no	google	eu	vi	**que**	é	um	é	um	é
e	o	amerício	eu	descobri	**que**	é	trecho_ingles.	agora	tô	olhando

Fonte: Elaborado para fins deste estudo. Destaque meu.

O Quadro 5 apresenta três linhas de concordância dos tradutores do termo “que”, que foram selecionadas aleatoriamente dentre as linhas com verbos realizando processos mentais com projeção. Os processos mentais (“sabia”, “vi” e “descobri”) encontram-se a uma posição à esquerda (coluna A5) do termo de busca.

Analisando a ocorrência dos PROCESSOS mentais (“sabia”, “vi” e “descobri”), todos eles apresentam projeção, caracterizando-se como metáforas interpessoais (como “sabia que” e “descobri que”). Pode-se dizer que os processos mentais, assim como os que caracterizam metáforas interpessoais, indicam metarreflexão por meio dos sujeitos, o que pode ser comprovado pela recorrência de “trecho_ingles” co-ocorrendo com o PARTICIPANTE “eu” à direita e “eu” co-ocorrendo à direita e à esquerda, como observado no Quadro 4.

[73] Segundo Halliday e Matthiessen (2014), metáfora interpessoal pode ser definida como uma realização metafórica (um item operando “como se fosse” de outro tipo) de um significado apresentando variação no sistema de MODO ou de MODALIDADE. Por exemplo, “acho que” apresenta um significado de probabilidade, uma vez que geralmente pode ser substituído por “possivelmente” ou “provavelmente”.

3.2.1.4 Listas de colocados

Tendo o mesmo objetivo da geração de listas de concordância – ou seja, obter evidências de metarreflexão dos sujeitos dos grupos dos pesquisadores e dos tradutores por meio da análise dos PROCESSOS mentais associados aos pronomes "eu" e "a gente" na função de SUJEITO – foram geradas também listas de colocados. Essas listas de colocados, utilizadas para obter os itens mais frequentes à esquerda e à direita desses pronomes em análise, podem, assim como as listas de concordância, fornecer evidências de metarreflexão nos textos dos pesquisadores e dos tradutores.

A fim de contabilizar tanto os PROCESSOS mentais com PARTICIPANTES realizados quanto os PROCESSOS mentais com PARTICIPANTES não realizados, foram realizados dois tipos de busca, com termos de busca distintos. No primeiro tipo de busca, visando os PROCESSOS mentais com PARTICIPANTES realizados, os termos buscados foram "eu" e "a gente", levando em conta os colocados à direita. Enquanto que no segundo tipo de busca, enfocando os PROCESSOS mentais com PARTICIPANTES não realizados, o termo de busca foi "que", considerando os colocados à esquerda. Essas escolhas se justificam pelo fato de que, segundo Halliday e Matthiessen (2014), os PROCESSOS em geral ocorrem com maior frequência à direita do PARTICIPANTE na estrutura da oração e, no caso de projeções, o PROCESSO (mental ou verbal) ocorre mais frequentemente à esquerda do "que".

Por isso, utilizando outra função do *R* elaborada, foram geradas as listas de colocados (à esquerda e à direita) dos termos "eu" e "a gente", considerando os itens que co-ocorrem até 5 termos à esquerda e 5 à direita. Foram obtidos, assim, os termos colocados à esquerda e à direita ordenados por sua frequência de modo decrescente. As listas de colocados à esquerda possuem os 20 termos mais frequentes e as linhas de colocados à direita os 20 termos mais relevantes (considerando a análise do "eu" e "a gente" relacionados aos PROCESSOS). As listas à direita foram analisadas de acordo com o tipo de PROCESSO, pois os verbos (que realizam os PROCESSOS) ocorrem com maior frequência à direita dos pronomes (que realizam os PARTICIPANTES). Por exemplo, na maioria dos textos, exceto talvez na ficção, "eu penso" é mais frequente que "penso eu".

Assim como foi feito anteriormente na apresentação das listas de frequência da Tabela 2 e da Tabela 3, são mostradas para cada grupo (pesquisadores e tradutores), as listas de colocados dos pronomes "eu", de "a gente" e dos processos mentais indicando metarreflexão

no relato dos sujeitos dos experimentos a respeito dos procedimentos realizados por eles durante a tarefa de tradução. As listas de colocados à direita do pronome "eu" são apresentadas na Tabela 8 e na Tabela 9 as listas de colocados à direita de "a gente" na Tabela 10 e na Tabela 11 e as listas de colocados à esquerda de "que" na Tabela 12 e na Tabela 13.

Tabela 8 – Lista dos vinte colocados relevantes mais frequentes de "eu" à direita para o grupo dos pesquisadores

Termo	**Frequência**
não	76
é	19
tive	18
acho	15
falei	12
pus	10
achei	9
fui	9
sabia	9
sei	9
vou	7
comecei	6
era	6
fiquei	6
tá	6
traduzi	6
achava	5
botei	5
coloquei	5
consegui	5

Fonte: Elaborada para fins deste estudo. Destaque meu.

Tabela 9 – Lista dos vinte colocados relevantes mais frequentes de "eu" à direita para o grupo dos tradutores

Termo	**Frequência**
não	37
fui	12
coloquei	11
tinha	11
tava	10
vou	10
fiquei	9
acho	8
deixei	7
fiz	7
traduzi	7
consertei	6
tô	5
troquei	5
achei	4
comecei	4
consultei	4
ficou	4
localizei	4
prefiro	4

Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Tabela 8 e na Tabela 9, podem ser observados os 20 colocados à direita mais relevantes (com base na limpeza de cada lista de colocados) para o grupo dos pesquisadores e dos tradutores do pronome "eu" organizados por frequência de forma decrescente. Devido à limpeza realizada, os termos além de "não" (marcando uma POLARIDADE negativa) são verbos, visando complementar a análise dos PROCESSOS mentais (como "acho" e "achei" na Tabela 8 e "achei" e "prefiro" na Tabela 9.

Observa-se também que os valores de frequência absoluta dos itens mais frequentes, como o "não", são maiores para o grupo dos pesquisadores, provavelmente devido à diferente no número de *tokens* nos textos de cada grupo (5383 para os pesquisadores e 2820 para os tradutores). Isso é corroborado pela observação da frequência relativa em relação ao número de *tokens* de cada texto, cujos valores são 1,41% (76/5383) para o grupo dos pesquisadores e 1,31% (37/2820) para o dos tradutores.

Observando os PROCESSOS mentais na Tabela 7 e na Tabela 8, notou-se uma maior ocorrência nos colocados dos pesquisadores (5 ocorrências – "acho", "achei", "sabia", "sei", "achava") que dos tradutores (3 ocorrências – "acho", "achei", "prefiro"). Isso pode ser devido tanto ao maior número de *tokens* nos textos dos pesquisadores quanto à maior variedade lexical de PROCESSOS mentais nos textos dos pesquisadores.

Tabela 10 – Lista dos treze colocados relevantes mais frequentes de "a gente" à direita para o grupo dos pesquisadores

Termo	**Frequência**
não	7
conhece	3
é	2
demora	1
faz	1
fica	1
gasta	1
pode	1
sabe	1
tem	1
usa	1
vai	1
vi	1

Fonte: Elaborada para fins deste estudo. Destaque meu.

Tabela 11 – Lista dos cinco colocados relevantes mais frequentes de "a gente" à direita para o grupo dos tradutores

Termo	**Frequência**
tem	2
é	1
não	1
tá	1
vacila	1

Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Tabela 10 e na Tabela 11, são apresentados, respectivamente, os 13 e os 5 colocados à direita mais relevantes (com base na limpeza de cada lista de colocados) para o grupo dos pesquisadores e dos tradutores do pronome "a gente" organizados por frequência de forma decrescente. Devido à limpeza realizada, além do termo "não" (marcando POLARIDADE

negativa), foram selecionados os verbos realizado PROCESSOS, visando complementar a análise dos PROCESSOS mentais (como “conhece” e “sabe” na Tabela 10).

Em seguida, é apresentada a análise dos processos mentais não realizados na Tabela 12 e na Tabela 13, que mostram as listas dos 20 colocados à esquerda relevantes mais frequentes de “que” para os textos dos pesquisadores e dos tradutores.

Tabela 12 – Lista dos vinte colocados relevantes mais frequentes de “que” à esquerda para o grupo dos pesquisadores

Termo	**Frequência**
é	46
não	38
acho	28
tem	24
foi	13
achei	9
sei	8
vi	7
tive	7
era	6
teria	6
lembrei	5
sabia	5
pode	4
conhece	3
fiquei	3
são	3
tinha	3
acha	2
achava	2

Fonte: Elaborada para fins deste estudo. Destaque meu.

Tabela 13 – Lista dos vinte colocados relevantes mais frequentes de “que” à esquerda para o grupo dos tradutores

Termo	**Frequência**
acho	11
não	7
foi	6
deixei	5
tem	4
deixa	3
tenho	3
tinha	3
achei	2
prefiro	2
pode	2
sei	2
tô	2
utilizei	2
vi	2
descobri	1
entendi	1
gostou	1
odiei	1
acho	11

Fonte: Elaborada para fins deste estudo. Destaque meu.

Na Tabela 12 e na Tabela 13, são apresentados os 20 colocados de “que” à esquerda mais relevantes (após a limpeza de cada lista de colocados) para o grupo dos pesquisadores e dos tradutores organizados por frequência de forma decrescente. Assim como nas tabelas anteriores, além do termo “não” (que marca POLARIDADE negativa), foram selecionados os

verbos realizando PROCESSOS, visando complementar a análise dos PROCESSOS mentais (como “acho” e “lembrei” na Tabela 12 e “acho” e “prefiro” na Tabela 13).

Da mesma forma que na Tabela 8, pode-se associar os PARTICIPANTES “eu” e “a gente” aos tipos de PROCESSO com os quais co-ocorrem, como “conhece”, “fica”, “gasta” e “vacila”, indicando metarreflexão no relato dos sujeitos dos experimentos a respeito dos procedimentos realizados por eles durante a tarefa de tradução.

Por fim, a ocorrência de processos mentais nas listas de colocados da Tabela 8, , da Tabela 9, da Tabela 10, da Tabela 12 e da Tabela 13 parece evidenciar a metarreflexão dos sujeitos de ambos os grupos. Isso provavelmente se deve aos PARTICIPANTES “eu” e “a gente” estarem associados aos PROCESSOS mentais utilizados e também possuírem a função de EXPERIENCIADOR, definido como o PARTICIPANTE que experiencia o significado do PROCESSO mental.

3.2.1.5 Considerações finais dos resultados dos dados não estruturados

Com base nos resultados apresentados na análise dos dados estruturados do tipo texto, conclui-se que esse tipo de dado permite a realização de uma grande gama de análises, como a análise dos pronomes “eu” e “a gente”. Além disso, utilizando o resultado (parcial ou total) de uma ferramenta como dado inicial para outra ferramenta, é possível chegar a conclusões ainda mais relevantes. Uma possibilidade é utilizar as orações das linhas de concordância para fazer uma lista de frequência e, em seguida, uma nuvem de palavras.[74] Com isso, é possível estudar os trechos de interesse da pesquisa selecionados nas linhas de concordância.

Como conclusão, PROCESSOS materiais e relacionais estão relacionados à representação de atividades de fazer e atribuir, seguidos dos mentais (incluindo instâncias de metáforas interpessoais) e verbais (relacionados à metarreflexão nas verbalizações dos sujeitos) (ALVES, 2003) os tipos de PROCESSO mental e as metáforas interpessoais (como “acho que” e “creio que”) indicam metarreflexão no discurso dos sujeitos dos experimentos ao relatar os procedimentos por eles realizados durante a tarefa de tradução. Isso pode ser explicado pelo fato de os PARTICIPANTES “eu” e “a gente” estarem frequentemente associados aos PROCESSOS

[74] Ferreguetti, em uma comunicação pessoal, constatou que isso era possível utilizando os *scripts* (ainda em uma versão não definitiva) da nuvem de palavras (que também gera uma lista de frequência) e das linhas de concordância.

mentais utilizados e possuírem, ao mesmo tempo, a função de EXPERIENCIADOR (o PARTICIPANTE que experiencia o PROCESSO mental). Essa conclusão corrobora Magalhães e Alves (2006), que afirmam que os processos mentais "mostram um equilíbrio entre processos que representam a busca a recursos internos" e "demonstram cognição de caráter mais deliberado, apontando para o desenvolvimento da capacidade de escolha e decisão dos tradutores" (MAGALHÃES; ALVES, 2006, p. 95).

3.2.2 Dados estruturados

3.2.2.1 Sumarização dos dados

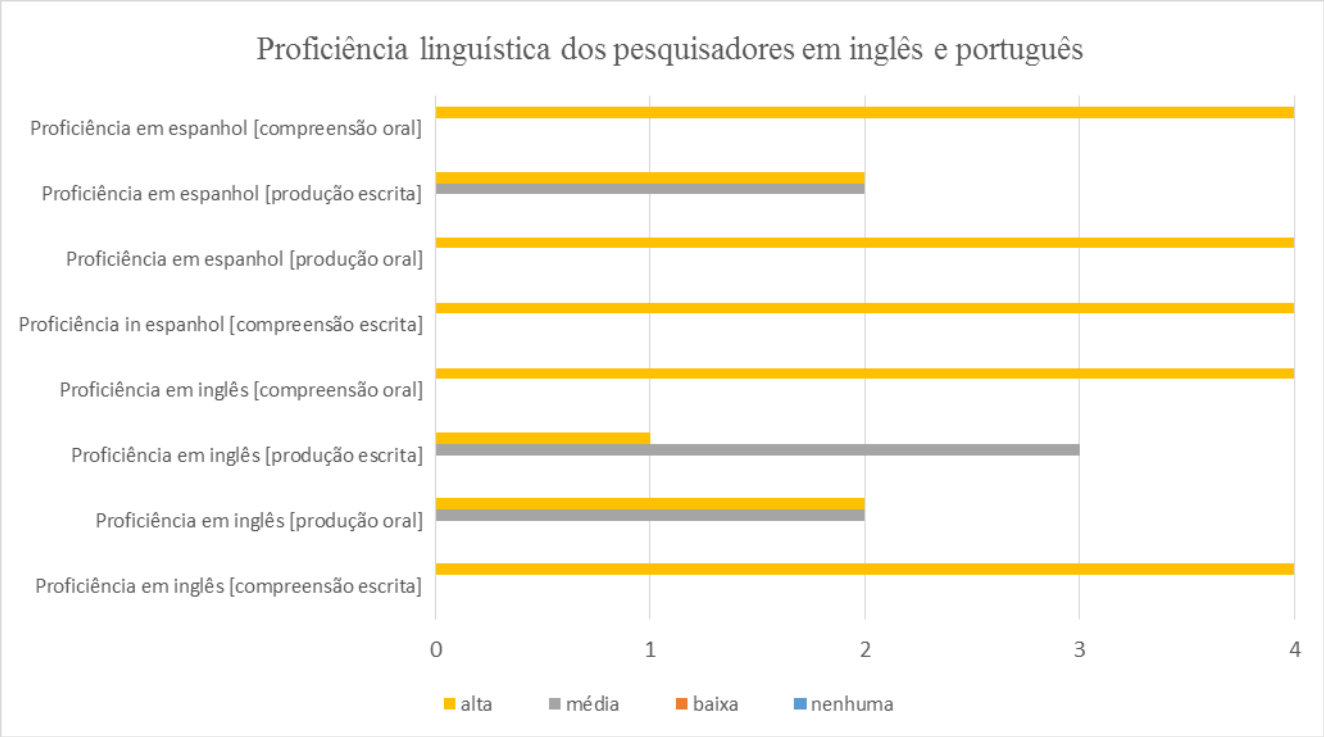
Por meio de *scripts* no *R*, utilizando funções nativas desse *software*, como a função *summary*, ou de pacotes, como a função *describeBy*, do pacote estatístico *psych*, foram sumarizados os dados dos questionários utilizados na análise dos dados estruturados. Nas seções a seguir são apresentados os dados resumidos dos pesquisadores (R1 a R4) e dos tradutores (T1 a T4) e alguns destes resultados são apresentados a seguir, nos gráficos, tabelas e quadros.

3.2.2.1.1 Dados dos pesquisadores

Utilizando os dados de proficiência linguística em inglês e espanhol das planilhas dos pesquisadores, proficiências que todos os pesquisadores sujeitos do estudo demonstraram possuir, foi elaborado o Gráfico 1.[75]

[75] Uma vez que as categorias e as respostas da planilha analisada nesta dissertação se encontram em inglês, os nomes das categorias serão apresentados em inglês nos gráficos, quadros e tabelas.

Gráfico 1 – Proficiência linguística dos pesquisadores em inglês e espanhol



Fonte: Elaborado para fins deste estudo.

O Gráfico 1 mostra a autoavaliação da proficiência linguística dos pesquisadores, considerando apenas a língua inglesa e a língua espanhola, das quais todos os sujeitos relataram possuir conhecimento. O eixo horizontal (com números de 0 a 4) mostra quantos dos quatro pesquisadores marcaram cada opção (relatadas pela legenda, da esquerda para a direita: “alta”, “média”, “baixa” e “nenhuma”) nos subitens das questões do questionário (distribuídos no eixo vertical). Por exemplo, enquanto no quesito “Proficiência em espanhol [produção escrita]” 2 dos 4 pesquisadores afirmaram ter proficiência alta e os outros dois, proficiência média, no quesito “Proficiência em inglês [produção escrita]” 1 sujeito afirmou possuir proficiência linguística alta, diferentemente dos demais, que se autoavaliaram possuindo proficiência média.

No Quadro 6, são apresentadas algumas estatísticas descritivas de categorias quantitativas relacionadas a local de moradia e informações sobre leitura e escrita em L1, L2 e L3.

Quadro 6 – Estatística descritiva de categorias quantitativas de local de moradia e informações sobre leitura e escrita em L1, L2 e L3 dos pesquisadores

Categoria quantitativa	Valor mínimo	Primeiro quartil	Mediana	Média	Terceiro quartil	Valor máximo
Número de meses de moradia em país falante de inglês	0	0	3,50	3,75	7,25	8,00
Número de meses de moradia em outro país	0	0	0	4,50	4,50	18,00
Número de textos diferentes lidos em L1	3,00	3,50	3,50	3,50	4,00	4,00
Número de textos diferentes lidos em L2	1,00	1,00	1,00	1,50	1,50	3,00
Número de textos diferentes lidos em L3	1,00	1,75	2,00	2,25	2,50	4,00
Número de textos diferentes escritos em L1	1,00	1,00	1,50	1,75	2,25	3,00
Número de textos diferentes escritos em L2	0	0,75	1,50	1,25	2,00	2,00
Número de textos diferentes escritos em L3	0	0,75	1,00	1,00	1,25	2,00

Fonte: Elaborado para fins deste estudo.

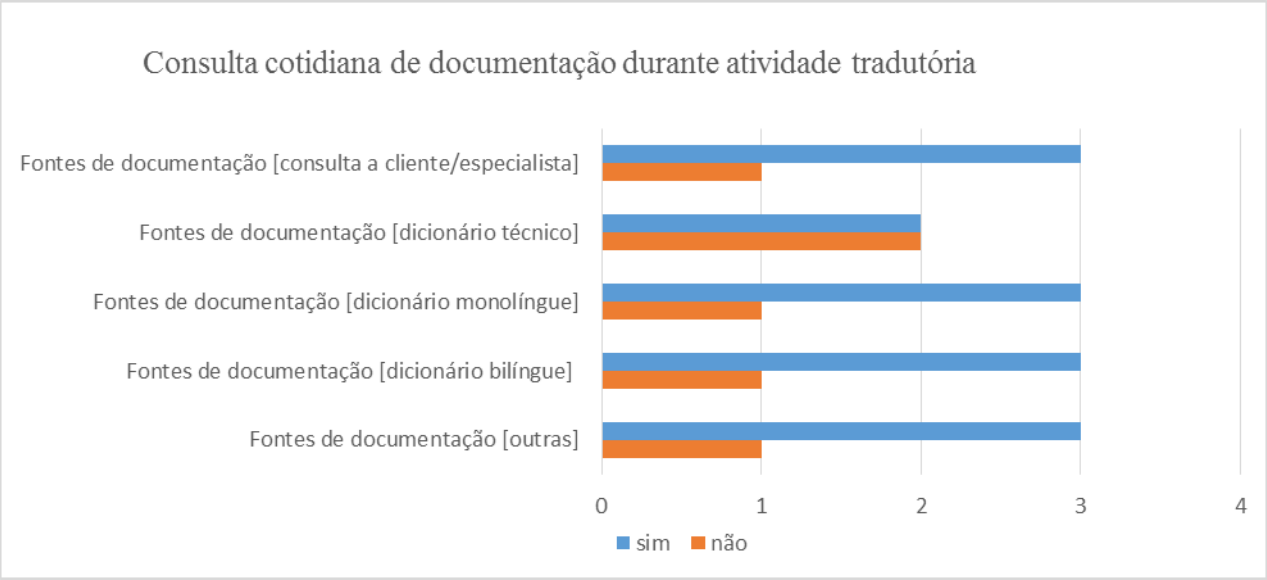
No Quadro 6 são apresentadas as estatísticas descritivas do valor mínimo, do primeiro quartil,[76] da mediana (ou segundo quartil), da média, do terceiro quartil e do valor máximo. Com essas informações é possível observar como os dados de interesse se distribuem e em que perfil se encaixam os sujeitos (caso esse seja o interesse), como no caso dos experimentos realizados. Além disso, utilizando outras medidas como o desvio padrão,[77] é possível verificar estatisticamente se cada sujeito apresentou valores típicos ou atípicos de seu grupo.

[76] O *quartil* é uma medida estatística que mede a classificação percentual de determinado dado considerando que um conjunto de dados foi ordenado e colocado em uma escala percentual com quatro pontos de referência, além do ponto inicial (o valor mínimo, equivalente a 0%): o primeiro quartil equivale a 0%, o segundo a 25% (também chamado de mediana), o terceiro a 50% e o quarto a 75% e o último (o valor máximo) é relativo a 100%.

[77] Na estatística, o desvio padrão representa o desvio (erro) de uma medição em relação à média, visto que toda medida se encontra no centro de um intervalo, com valores abaixo e acima do valor central (KERNS, 2011, p. 38).

Por fim, no Gráfico 2, podem-se observar os hábitos de consulta a fontes de apoio para a tradução relatados pelos pesquisadores.

Gráfico 2 – Consulta cotidiana de documentação de pesquisa dos pesquisadores durante atividade tradutória



Fonte: Elaborado para fins deste estudo.

No Gráfico 2, é possível observar que todos os tradutores utilizam todas as opções de documentação para pesquisa, ou seja, consulta ao cliente/especialista, dicionário técnico, dicionário monolíngue, dicionário bilíngue e outras fontes de pesquisa (outras), não citadas no questionário. Assim como no Gráfico 1, o eixo horizontal mostra quantos dos 4 pesquisadores marcaram cada opção (“sim” e “não”) nos subitens das questões do questionário envolvidas (distribuídos no eixo vertical).

3.2.2.1.1 Dados dos tradutores

No Quadro 7, a seguir, são apresentadas algumas estatísticas descritivas de categorias quantitativas relacionadas a local de moradia em país falante de inglês e de rendimento ao fazer traduções inversas (no caso, inglês-português brasileiro) e diretas (no caso, português brasileiro-inglês). Apresentam-se também informações sobre leitura e escrita em L1, L2 e L3.

Quadro 7 – Estatística descritiva de categorias quantitativas de moradia em país falante de inglês e de rendimento ao fazer traduções inversas e diretas dos tradutores

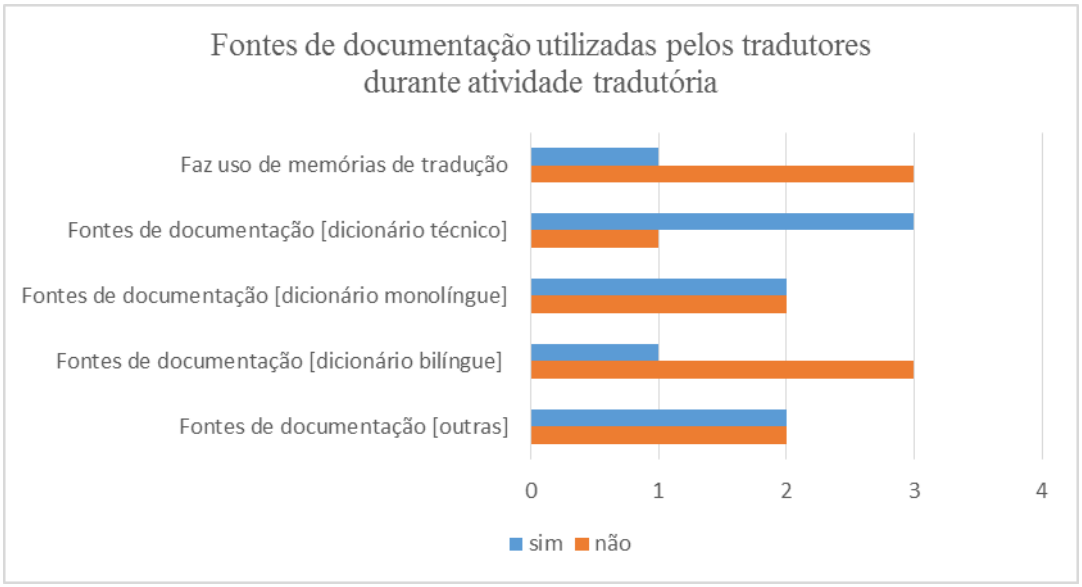
Categoria quantitativa	Valor mínimo	Primeiro quartil	Mediana	Média	Terceiro quartil	Valor máximo
Número de meses de moradia em país falante de inglês	0	0	24,00	33,00	57,00	84,00
Número médio de páginas traduzidas diariamente do português para o inglês nos últimos 2 anos	2,00	2,00	2,00	2,33	2,50	3,00
Número médio de páginas traduzidas diariamente do inglês para o português nos últimos 2 anos	3,00	3,00	3,00	3,00	3,00	3,00

Fonte: Elaborado para fins deste estudo.

No Quadro 7 são apresentadas as estatísticas descritivas do valor mínimo, do primeiro quartil, da mediana, da média, do terceiro quartil e do valor máximo de algumas categorias para os tradutores, assim como o Quadro 6 apresenta as mesmas características para os pesquisadores (em categorias diferentes, devido às diferenças dos questionários). Com essas informações é possível traçar o perfil dos sujeitos a partir de padrões presentes nas respostas dos questionários. Por exemplo, o fato de que, para a categoria "número médio de páginas traduzidas diariamente do inglês para o português nos últimos dois anos" (*Average number of pages translated daily from English into Portuguese in the last 2 years*), o valor mínimo, a mediana, a média e o valor máximo são os mesmos indica que os quatro sujeitos responderam "3" para esse item, sem demonstrar nenhuma variação em suas respostas.

A seguir, o Gráfico 3 apresenta informações sobre os hábitos de consulta de documentação de pesquisa durante atividade tradutória dos tradutores.

Gráfico 3 – Consulta cotidiana de documentação de pesquisa dos tradutores durante atividade tradutória



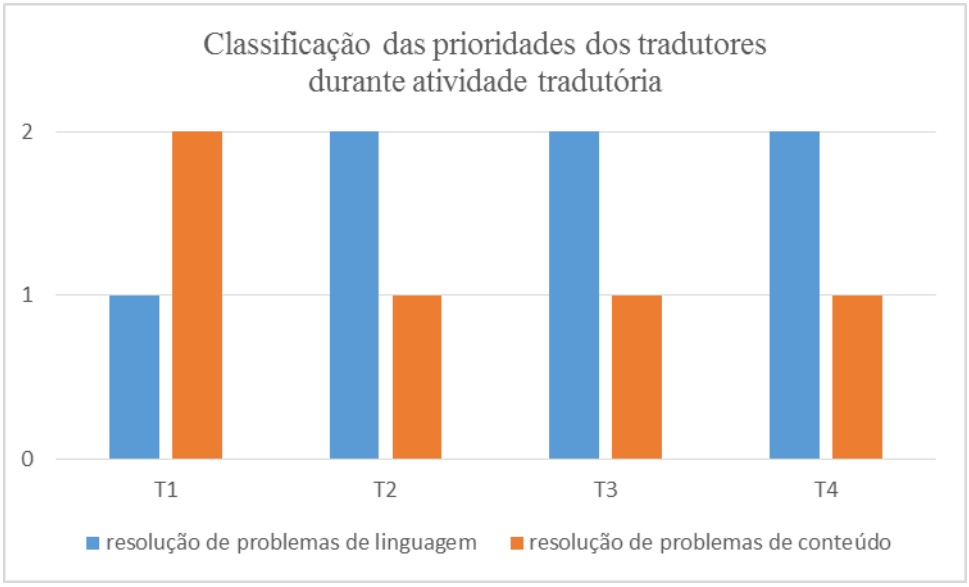
Fonte: Elaborado para fins deste estudo.

No Gráfico 4, que apresenta a resposta dos tradutores sobre seus hábitos de consulta ao realizar seu trabalho, assim como no Gráfico 2, todos os tradutores relataram utilizar as opções de documentação de pesquisa,[78] embora provavelmente nem todos as utilizem simultaneamente. Essas opções são: memórias de tradução, dicionário técnico, dicionário monolíngue, dicionário bilíngue e outras fontes de pesquisa, não citadas no questionário.

Nos próximos três gráficos, o Gráfico 4, o Gráfico 5 e o Gráfico 6, são apresentadas as classificações realizadas pelos tradutores (T1 a T4) das prioridades de resolução de problemas, comparando sempre duas opções por vez. Se o sujeito marcou “1” como a prioridade para determinado item, esse item é mais importante que o outro (marcado como “2”). Por exemplo, para o sujeito T2, resolver problemas de conteúdo é mais importante que resolver problemas de linguagem.

[78] Uma vez que a investigação das fontes de documentação utilizadas por tradutores em pesquisas dos Estudos da Tradução se encontram além do escopo desta dissertação, essas opções foram utilizadas como categoria de análise dos dados coletados.

Gráfico 4 – Classificação das prioridades dos tradutores entre duas opções de resolução de problemas (de linguagem ou de conteúdo)

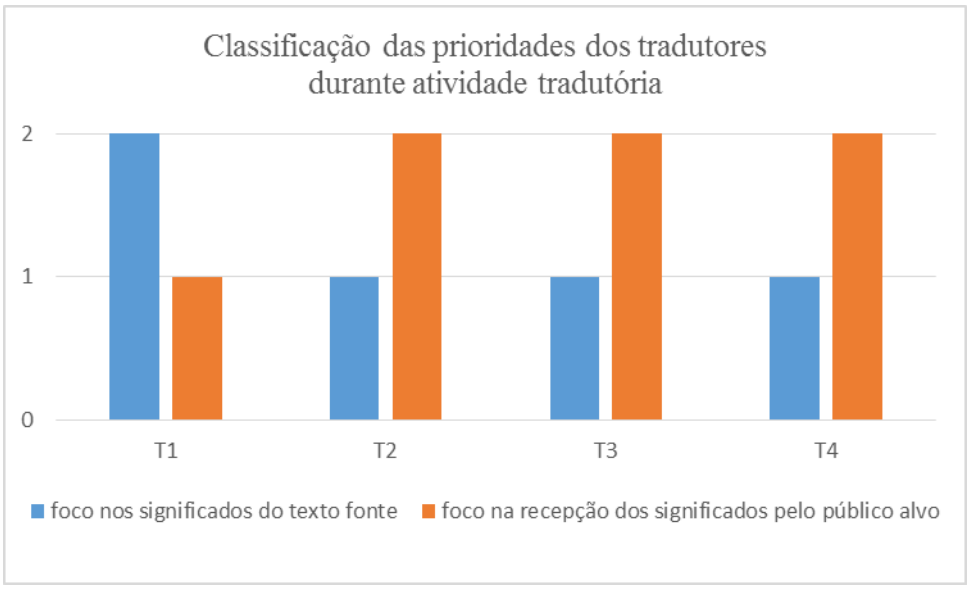


Fonte: Elaborado para fins deste estudo.

O Gráfico 4 mostra a preferência dos tradutores entre a resolução de problemas de linguagem (envolvendo questões linguísticas) e de conteúdo (envolvendo principalmente questões lexicais). O primeiro tradutor (T1), ao contrário dos demais, relata ser mais importante resolver os problemas de linguagem que os de conteúdo.

O Gráfico 5, a seguir, mostra a preferência dos tradutores entre se aproximar do significado do texto-fonte ou do texto-alvo.

Gráfico 5 – Classificação das prioridades dos tradutores entre a aproximação do texto-fonte ou do texto-alvo

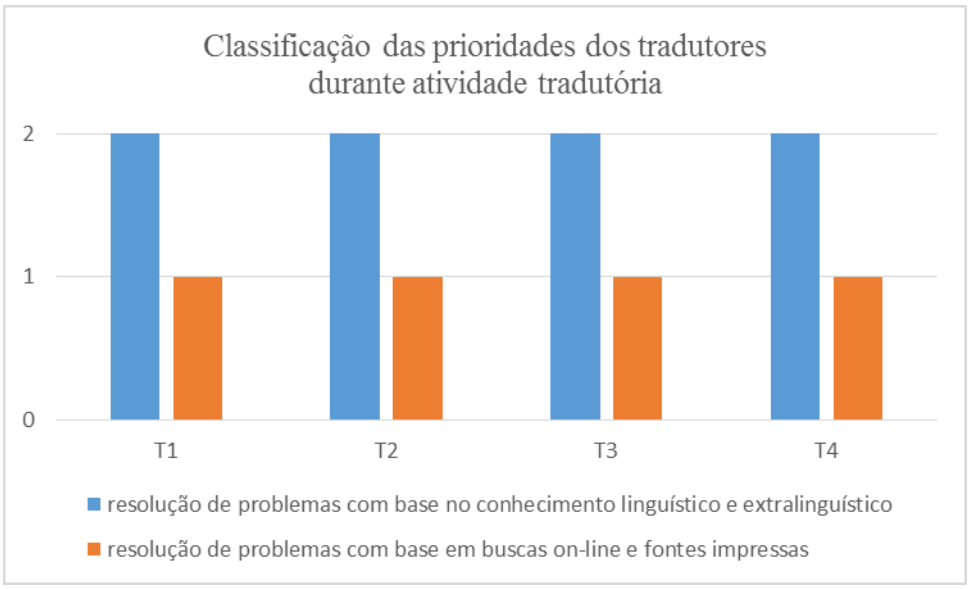


Fonte: Elaborado para fins deste estudo.

Pode-se observar no Gráfico 5 a preferência dos tradutores entre se aproximar do significado do texto-fonte ou do texto-alvo. Da mesma forma que na Figura 5, o primeiro tradutor (T1) se comportou diferentemente dos demais, dando mais importância ao significado do texto-alvo que do texto-fonte. Isso relacionado aos conceitos de "domesticação" e "estrangeirização" dos Estudos da Tradução, que não serão discutidos pois excedem o escopo deste trabalho (cf. VENUTI, 2000).

O Gráfico 6 apresenta a preferência dos tradutores entre fazer uso do próprio conhecimento linguístico ou da consulta de fontes externas ao traduzir.

Gráfico 6 – Classificação das prioridades dos tradutores entre a resolução de problemas por meio do próprio conhecimento linguístico ou pela consulta de fontes externas



Fonte: Elaborado para fins deste estudo.

No Gráfico 6, que apresenta a preferência dos tradutores entre o uso do próprio conhecimento linguístico ou da consulta de fontes externas ao traduzir, pode-se notar um padrão diferente do ocorrido no Gráfico 4 e no Gráfico 5. No Gráfico 6, O novo padrão mostra que todos os tradutores se comportaram da mesma forma, indicando preferência pela resolução de problemas pela consulta de fontes externas, não de seu próprio conhecimento.

3.2.2.2 Agrupamento dos sujeitos por meio de dendrogramas

Por meio do dendrograma, a ferramenta de análise utilizada nesta dissertação para lidar com dados estruturados, é possível agrupar os sujeitos de acordo com os dados dos questionários (como já citado, de um tipo para os pesquisadores e de outro para os tradutores). Para realizar essa tarefa, os dados foram importados e adequados para a análise, sendo essencial a conversão, via *R*, dos dados qualitativos (referentes às categorias) em *fatores*[79] para que eles pudessem ser estruturados em matrizes numéricas. Essas matrizes, por sua vez, se organizam em linhas e colunas. No caso desta pesquisa, as linhas apresentam os sujeitos, codificados como

[79] *Fator* é um tipo de dado do *R* que é utilizado para processar categorias, que são contadas a fim de descrever sua distribuição.

R1 a R4 (os pesquisadores) e como T1 a T4 (os tradutores). As colunas, por sua vez, apresentam as categorias de análise.

A partir de esses dados, foram criados dendrogramas utilizando a função elaborada para tal fim, que utilizou a distância euclidiana,[80] ou distância entre pontos no gráfico cartesiano, e o método Ward[81] para fazer a separação entre os sujeitos. Foram feitos três tipos de agrupamentos entre os sujeitos comparando seus dados: (i) uma comparação dos 4 pesquisadores; (ii) uma comparação entre os 4 tradutores; (iii) uma comparação entre todos os sujeitos, utilizando apenas os dados fornecidos pelas perguntas em comum entre ambos os tipos de questionários., os quais se referiam às categorias: "Experiência de moradia em país de língua inglesa", "Fontes de documentação [dicionário bilíngue]", "Fontes de documentação [dicionário monolíngue]", "Fontes de documentação [dicionário técnico]", "Fontes de documentação [outras]"

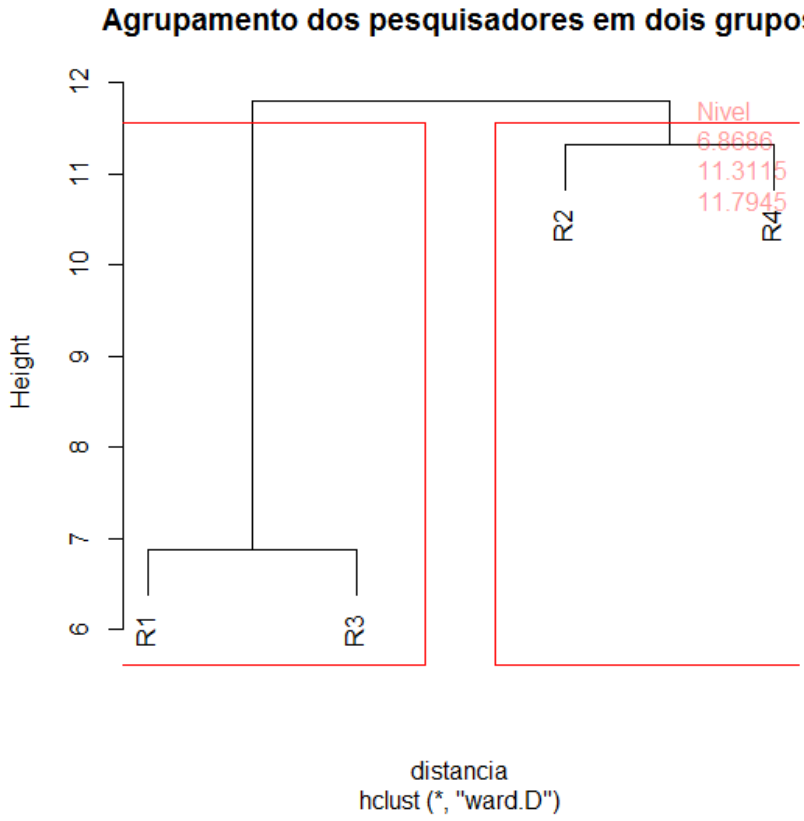
Os dendrogramas relativos a esses agrupamentos são apresentados e descritos nas subseções a seguir, no Gráfico 7, no Gráfico 8, no Gráfico 9 e Gráfico 10.

[80] Segundo Linden (2009), a distância euclidiana é uma medida simples de ser calculada, utilizada quando não é necessária uma sensitividade maior no agrupamento quanto às maiores distâncias entre os elementos.
[81] Segundo Hair *et al*. (2005), o método de Ward consiste em uma forma de agrupar hierarquicamente elementos utilizando como medida de similaridade as distâncias entre os dois agrupamentos sobre todas as variáveis (categorias) analisadas. Com esse método, é possível diminuir a variabilidade entre os agrupamentos e diminuir a influência dos valores extremos nas distâncias.

3.2.2.2.1.1 Comparação dos dados dos pesquisadores

Essa seção apresenta a comparação dos dados dos pesquisadores por meio do dendrograma do Gráfico 7.

Gráfico 7 – Agrupamento dos pesquisadores de acordo com as categorias dos questionários



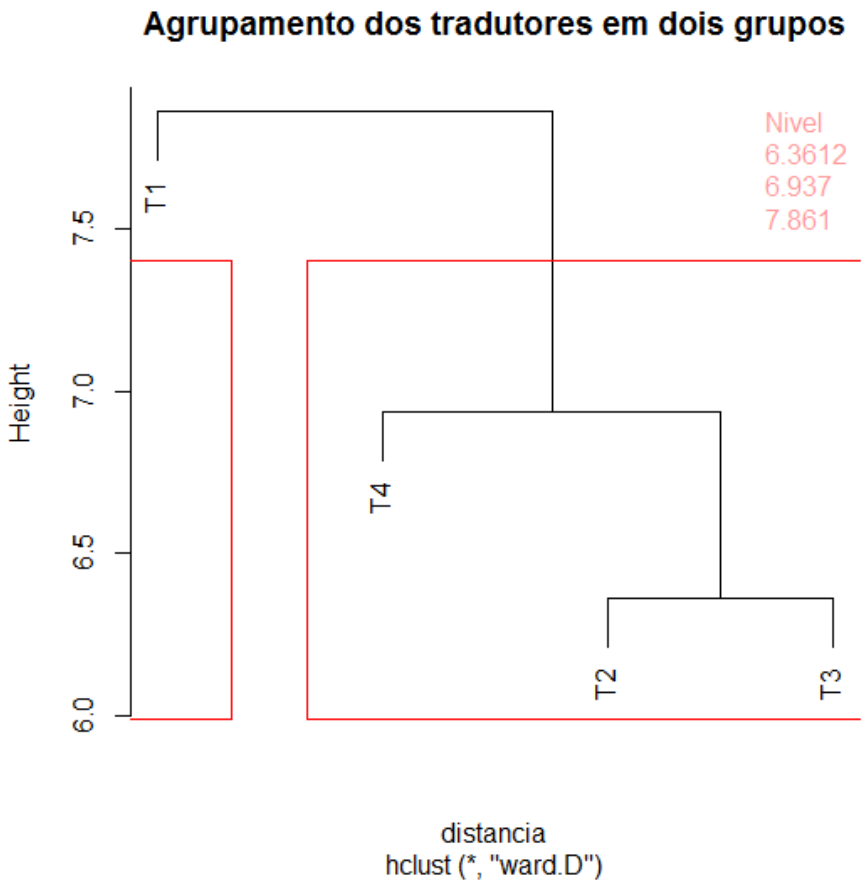
Fonte: Elaborado para fins deste estudo.

Como visto no Gráfico 7, os pesquisadores, identificados pelos códigos R1 a R4, foram divididos em dois grupos: R1 e R3; R2 e R4. A separação desses grupos fica clara devido à divisão definida pelas linhas vermelhas e no canto superior direito é possível observar as distâncias (medidas em relação ao eixo vertical) relativas aos níveis de cada grupo. Essas distâncias, calculadas pelo *software* e apresentadas na escala do gráfico, representam uma medida de dissimilaridade entre os agrupamentos de textos.

3.2.2.2.1.2 Comparação dos dados dos tradutores

Essa seção apresenta a comparação dos dados dos tradutores por meio do dendrograma do Gráfico 8.

Gráfico 8 – Agrupamento dos tradutores de acordo com as categorias dos questionários



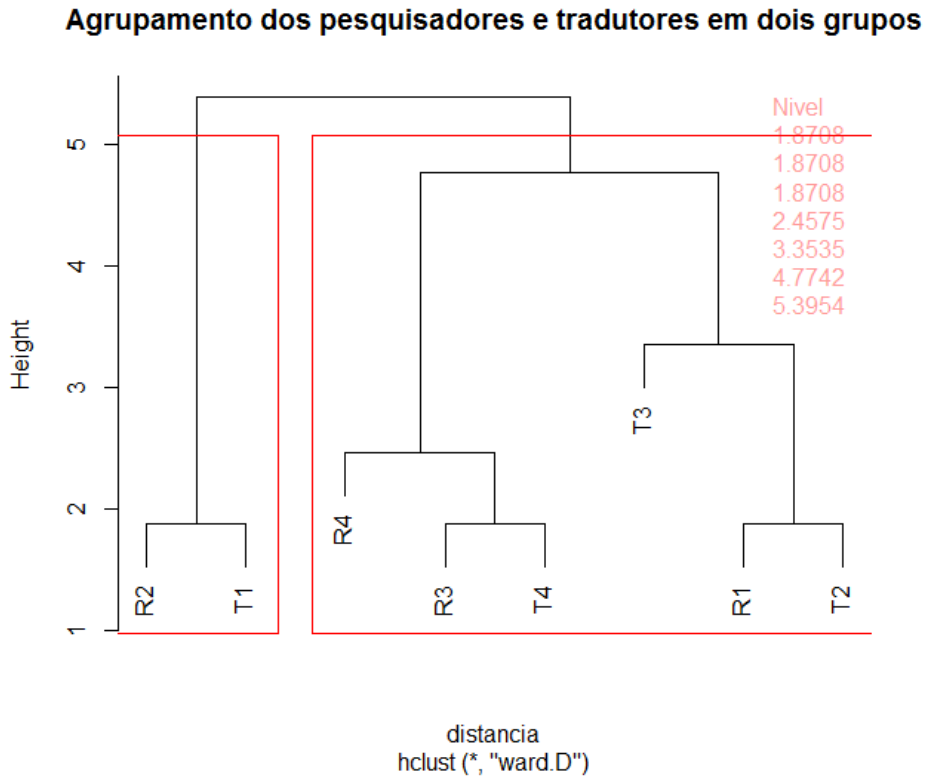
Fonte: Elaborado para fins deste estudo.

No Gráfico 8, os tradutores, identificados pelos códigos T1 a T4, foram divididos em dois grupos: T1; e T4, T2 e T3. Assim como no Gráfico 7, os grupos se separam pelas linhas vermelhas e no canto superior direito se encontram as distâncias (medidas em relação ao eixo vertical) relativas aos níveis de cada grupo. Com isso, o Gráfico 8 corrobora o que foi observado no Gráfico 4 e no Gráfico 5, nos quais T1 parece possuir um perfil diferenciado entre os tradutores.

3.2.2.2.1.3 Comparação dos dados comuns entre pesquisadores e tradutores

Para a comparação entre os pesquisadores e tradutores, foram utilizadas apenas as categorias comuns entre os sujeitos, especificamente: “Experiência de moradia em país de língua inglesa”, “Fontes de documentação [dicionário bilíngue]”, “Fontes de documentação [dicionário monolíngue]”, “Fontes de documentação [dicionário técnico]”, “Fontes de documentação [outras]”. Tanto para os pesquisadores (R1 a R4) quanto para os tradutores (T1 a T4), foram elaborados os dendrogramas a seguir, dividindo os sujeitos em, respectivamente, dois e três grupos, apresentados no Gráfico 9 e no Gráfico 10.

Gráfico 9 – Agrupamento dos pesquisadores e dos tradutores em dois grupos de acordo com as categorias comuns de ambos os questionários

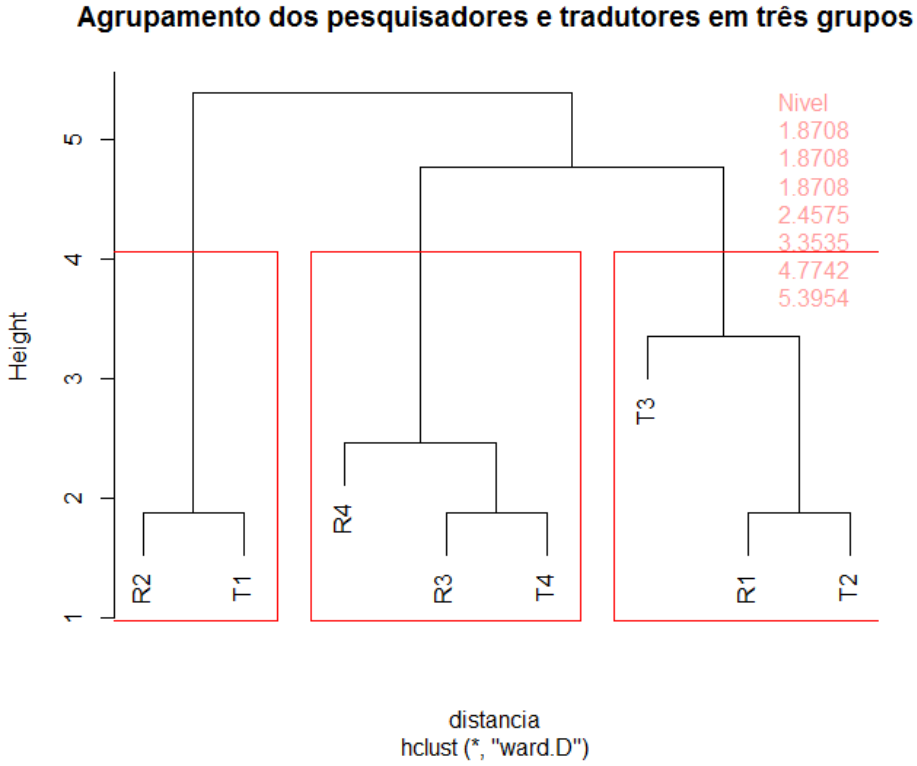


Fonte: Elaborado para fins deste estudo.

No Gráfico 9, os tradutores e os pesquisadores foram divididos em dois grupos: R2 e T1; R4, R3, T4, T3, R1 e T2. Como já mencionado, os grupos separados pelas linhas vermelhas e todas as distâncias entre os níveis de cada separação se encontram no canto superior direito, medidas em relação ao eixo vertical.

O Gráfico 10 que utiliza os mesmos dados que o Gráfico 9, apresenta o agrupamento dos pesquisadores e dos tradutores em três grupos com base nas categorias comuns de ambos os questionários.

Gráfico 10 – Agrupamento dos pesquisadores e dos tradutores em três grupos de acordo com as categorias comuns de ambos os questionários



Fonte: Elaborado para fins deste estudo.

No Gráfico 10, é possível observar que os tradutores e os pesquisadores se dividem agora em três grupos, e não em dois. Os três grupos, separados pelas linhas vermelhas, são: R2 e T1; R4, R3, T4; T3, R1 e T2. As distâncias entre os níveis de separação, no canto superior direito, são as mesmas do dendrograma do Gráfico 9, deixando claro que a principal diferença entre esses dois dendrogramas é o número de grupos, considerando que quanto maior o número de grupos maior a homogeneidade dentro dos grupos e maior a diferença entre os grupos.

3.2.2.3 Considerações finais dos resultados dos dados estruturados

Com base na sumarização dos dados estruturados por meio de gráficos e quadros, os quais puderam ser descritos por medidas de Estatística Descritiva como valor mínimo, valor máximo, média, mediana e o primeiro e o terceiro quartis, diversos padrões nos dados puderam ser observados.

Tendo em vista os dendrogramas apresentados anteriormente, é possível observar que, uma vez que os elementos a serem agrupados tenham seus dados preparados adequadamente, é possível agrupá-los por meio de dendrogramas. Além disso, comparando os dendrogramas do Gráfico 9 e do Gráfico 10, que foram elaborados tendo como base os mesmos dados, no primeiro caso há apenas 2 grupos e no segundo 3 grupos. Isso mostra que, se houver vários itens a serem agrupados, é possível continuar fazendo o agrupamento pela subdivisão dos grupos já existentes até certo ponto, desde que ainda haja elementos para compor os grupos. Esse agrupamento sucessivo pode influenciar na análise realizada pois quanto maior o número de grupos maior a homogeneidade dentro dos grupos e maior a diferença entre os grupos.

De fato, embora nesta implementação das ferramentas elaboradas tenha sido feita a divisão desses sujeitos em dois e três grupos, é possível dividi-los por meio do dendrograma em até sete grupos. Em outras palavras, pode-se separar os oito elementos em até sete grupos considerando os dados utilizados – o número de grupos depende principalmente da hipótese inicial ou do objetivo do pesquisador ao dividir esses sujeitos.

Além disso, deve-se destacar a triangulação de dados observada entre os dados estruturados dos tradutores sumarizados – nos gráficos da Gráfico 4 e da Gráfico 5 – e seu agrupamento no Gráfico 8, a qual indica que T1 parece possuir um perfil diferenciado entre os tradutores.

3.3 Considerações finais sobre os dados estruturados e não estruturados

Neste capítulo foi apresentada a análise dos dados não estruturados do tipo texto e dos dados estruturados, os quais foram descritos por medidas de Estatística Descritiva como valor mínimo, valor máximo, média, mediana e quartis e agrupados em dendrogramas.

A partir disso, foi possível observar a forma como informações foram obtidas dos textos utilizando a lista de frequência, a nuvem de palavras, as linhas de concordância e as listas de

colocados. Também foi possível sumarizar e agrupar os dados de questionários por meio de funções estatísticas do ambiente *R* e do uso do dendrograma.

Uma vez que um dos objetivos desta dissertação foi apresentar ferramentas de análise de dados estruturados e não estruturados, demonstrou-se também que é possível fazer uma análise de determinado objeto de estudo, seja no processo ou no produto da tradução, combinando as ferramentas de ambos os tipos de dados. A título de exemplo, caso o interesse de pesquisa fosse analisar por que razões o pesquisador R2 e o tradutor T2 foram mais bem avaliados (BRAGA, 2012), é possível chegar a resultados utilizando as ferramentas mostradas nesse capítulo para analisar dados estruturados e não estruturados. As evidências informadas nesta dissertação, principalmente com base nos dendrogramas utilizados na análise dos dados estruturados, indicam que R2 e T2 não se destacaram nos dendrogramas, o que sugere a necessidade de investigações posteriores. Esse tipo de pesquisa, por exemplo, poderia ser realizada tendo como base os protocolos verbais de R2 e T2 e os dados dos questionários desses sujeitos, aplicando essas ferramentas para analisar e tirar conclusões com base nos resultados gerados e utilizando como teoria linguística a Linguística Sistêmico-Funcional.

Em suma, os tipos de PROCESSO mental e as metáforas interpessoais (como “acho que” e “creio que”) indicam metarreflexão nos textos dos sujeitos dos experimentos de ambos os grupos. Isso se deve possivelmente ao fato de os PARTICIPANTES “eu” e “a gente” frequentemente estarem associados aos PROCESSOS mentais utilizados e possuírem, ao mesmo tempo, a função de EXPERIENCIADOR – o PARTICIPANTE que experiencia o significado do PROCESSO mental. Isso corrobora Magalhães e Alves (2006), pois os processos mentais “mostram um equilíbrio entre processos que representam a busca a recursos internos” e “demonstram cognição de caráter mais deliberado, apontando para o desenvolvimento da capacidade de escolha e decisão dos tradutores” (MAGALHÃES; ALVES, 2006, p. 95).

CONSIDERAÇÕES FINAIS

Afiliada aos Estudos da Tradução, no escopo da Linguística com potencial de aplicação (HALLIDAY, 1985), desenhada no marco teórico da Linguística Sistêmico-Funcional (HALLIDAY; MATTHIESSEN, 2014), esta pesquisa utiliza subsídios da Linguística de *Corpus*, da Mineração de dados e de textos, da Estatística Descritiva e de técnicas multivariadas de análise. Esta dissertação apresenta um estudo de desenvolvimento, implementação e teste de um conjunto de ferramentas de preparação e análise de dados estruturados (em planilhas) e não estruturados (do tipo texto), utilizando-se *scripts* elaborados no *software* estatístico e ambiente computacional *R*.

O capítulo 1 apresentou a revisão das principais teorias que informam esta dissertação, tais quais trabalhos dos Estudos da Tradução que fazem uso de métodos quantitativos, como Oakes e Ji (2012), os quais são contrastados com trabalhos que não utilizam tais métodos, como Baker (2000) e Malmkjaer (2004). Por fim, é apresentado o conceito de Linguística com potencial de aplicação (*Appliable Linguistics*), ao qual esta dissertação se afilia utilizando como fontes de subsídios conhecimentos de outras áreas, quais sejam, a Linguística de *Corpus*, a Mineração de dados e de textos e a Estatística Aplicada aos estudos linguísticos.

Levando em consideração Malmkjaer (2005), esta dissertação contribui para os Estudos da Tradução aplicando conhecimentos de outro(s) campo(s) disciplinar(es) (no caso, a Ciência da Computação e a Estatística) no estudo do fenômeno da tradução, permitindo que achados e percepções dessas áreas do conhecimento auxiliem na melhor compreensão desse fenômeno (MALMKJAER, 2005, p. 20-21). Mais especificamente, esta dissertação propõe ferramentas que possibilitam a análise de dados extraídos de textos traduzidos (produto) e dados gerados na execução de pesquisas do processo tradutório, gerados em condições experimentais.

Os dados utilizados são provenientes de experimentos realizados no LETRA com 4 pesquisadores do CDTN e 4 tradutores profissionais, que preencheram um formulário com dados sociodemográficos e outras informações, tais como hábitos de leitura e conhecimentos linguísticos. Posteriormente, foram analisados dados previamente registrados de protocolos verbais (livres e guiados) da tarefa realizada. Os dados utilizados nesta dissertação são, portanto, do tipo estruturado e não estruturado, respectivamente.

No capítulo 2 foram apresentados os procedimentos de preparação dos dados e os *softwares* utilizados para o processamento e a análise dos dados, quais sejam, o *Notepad++* e

o *R*, assim como a metodologia de análise de cada tipo de dado – não estruturado e estruturado – descritas em separado, pois para cada tipo são utilizados métodos diferentes, embora possuam semelhanças entre si (REZENDE, 2003). Com os dados não estruturados, foram obtidas listas de frequência, uma nuvem de palavras, linhas de concordância e listas de colocados a partir dos protocolos verbais. Com os dados estruturados, por sua vez, foi feita a sumarização (resumo) desses dados tendo como base a Estatística Descritiva (com o uso de média, mediana, valor mínimo, valor máximo, dentre outras medidas) e também o agrupamento dos sujeitos dentro dos grupos (pesquisadores e tradutores) e entre todas as amostras.

Esta dissertação cumpriu seu objetivo geral, de desenvolver, implementar e testar um conjunto de ferramentas de preparação e análise de dados estruturados e não estruturados para a realização de estatísticas aplicáveis a esses dados, explorando-se a potencialidade do ambiente *R*. Para isso, as seguintes etapas foram realizadas:

- elaboração dos *scripts* do ambiente *R* para auxiliar na preparação de dados estruturados e não estruturados;
- desenvolvimento de uma metodologia baseada na Mineração de dados e de textos com base em ferramentas da Linguística de Corpus para extrair os dados de planilhas de dados e de textos no ambiente *R*;
- implementação e teste dos *scripts* elaborados em uma amostra de transcrições de protocolos verbais de experimentos (dados não estruturados) e em um banco de dados de informações relativas aos tradutores envolvidas nesses experimentos (dados estruturados), com o objetivo de triangular os dados gerados e as estatísticas aplicadas à linguística;
- interpretação dos resultados da implementação das ferramentas elaboradas utilizando como subsídio a Linguística Sistêmico-Funcional, mais especificamente na análise dos PROCESSOS mentais (com participantes "eu" e "a gente" realizados ou não) sob a perspectiva discursiva como indicadores do posicionamento do sujeito do experimento nos protocolos verbais sobre a tarefa executada.

Como consequência do cumprimento dos objetivos, a implementação da metodologia de pesquisa elaborada permitiu a geração de resultados que puderam ser interpretados com subsídio da Linguística Sistêmico-Funcional.

Como pôde ser visto no capítulo 3, foi descrita a aplicação da metodologia de pesquisa de dados estruturados e não estruturados apresentada no capítulo 2, bem como foram discutidos os resultados gerados. Nessa discussão, são apontadas as conclusões e outras possibilidades de análise integrando o uso de dados estruturados e não estruturados em uma abordagem quali-quantitativa utilizando os métodos quantitativos para obter conclusões que derivem da complementaridade dos dados qualitativos e quantitativos analisados. Ou seja, fez-se uso de uma abordagem quantitativa com subsídio de ferramentas computacionais e estatísticas. Além disso, foi feita a triangulação dos resultados obtidos para os dados estruturados, especificamente entre os dados sumarizados e os dendrogramas, que agruparam os sujeitos dentro de cada grupo e entre todos de acordo com as planilhas de dados.

Os resultados obtidos para a análise dos dados foram gerados separadamente para os dados não estruturados e estruturados, tendo como enfoque os pronomes “eu” e “a gente”, realizações dos PARTICIPANTES dos protocolos transcritos, e os verbos com os quais co-ocorrem, realizações dos PROCESSOS, que explicam o comportamento dos PARTICIPANTES. Para enfocar os casos que não houve realização dos PARTICIPANTES, analisou-se também o “que”, que frequentemente é utilizado como parte de uma projeção do PROCESSO mental. O objetivo desta análise foi verificar evidências de metarreflexão dos sujeitos dos experimentos a partir dos PARTICIPANTES e PROCESSOS, com destaque aos PROCESSOS mentais.

Para os dados não estruturados, foram obtidas listas de frequência, uma nuvem de palavras, linhas de concordância e listas de colocados, a partir dos quais notou-se a ocorrência de PROCESSOS materiais, relacionais (associados à representação de atividades de fazer e atribuir), mentais (incluindo instâncias de metáforas interpessoais) e verbais (relacionados à metarreflexão nas verbalizações dos sujeitos) em co-ocorrência com os pronomes “eu” e “a gente” e com o “que”. Com isso, por meio da co-ocorrência desses itens com os PROCESSOS mentais e de trechos em inglês e em português do resumo do experimento, pôde-se concluir que houve evidência de metarreflexão dos sujeitos com base no relato das atividades durante a execução da atividade de tradução.

Para os dados estruturados, foram gerados dados sumarizados, os quais foram triangulados com o agrupamento dos sujeitos e a geração de dendrogramas para agrupar os sujeitos dentro dos grupos de pesquisadores e tradutores e entre todos. Com base nisso, visou-

se verificar que características separavam sujeitos que apresentaram comportamento atípico, isto é, agrupados separados dos demais nos dendrogramas.

Acerca dos PROCESSO mental e das metáforas interpessoais, ambos indicam metarreflexão nos textos dos sujeitos dos experimentos dos pesquisadores e dos tradutores, uma vez que os verbos na primeira e terceira pessoa do singular, quando possuem PARTICIPANTES realizados, utilizam "eu" e "a gente" frequentemente associados aos PROCESSOS mentais e operam, simultaneamente, na função de EXPERIENCIADOR (o PARTICIPANTE que experiencia o PROCESSO mental). Isso está de acordo com Magalhães e Alves (2006, p. 95), que afirmam que os PROCESSOS mentais "demonstram cognição de caráter mais deliberado, apontando para o desenvolvimento da capacidade de escolha e decisão dos tradutores".

CONTRIBUIÇÕES

Em suma, esta dissertação contribui não apenas para os Estudos da Tradução (para a análise do produto e do processo tradutório), como também para os estudos linguísticos em geral, visto que oferece uma nova metodologia de análise de dados estruturados em planilhas de dados e de dados não estruturados do tipo texto.

Por meio do uso dessa metodologia de análise, que utiliza como base o editor de textos *Notepad++* para a preparação dos dados e o *software* estatístico e ambiente computacional *R* tanto para a preparação quanto para a análise dos dados, é possível fazer diversos tipos de análise utilizando parte ou a totalidade das ferramentas disponibilizadas. Essas ferramentas são funções do *R* elaboradas e organizadas em um *script*, que consiste em arquivo de texto com formato *.R*, o qual pode ser importado por meio do ambiente *R* e permitir a geração dos resultados, assim como mostrado no capítulo 3.

Pelo ponto de vista do produto (o texto traduzido), pode-se utilizar a metodologia elaborada para realizar a análise de dados não estruturados (nesse caso, de textos) em um único *software*, o *R*. Isso é possível porque este *software* e ambiente de programação permite a realização de todas as etapas da pesquisa, desde a preparação dos dados, passando pela análise e a apresentação dos resultados.

Pelo ponto de vista do processo tradutório, essa metodologia pode ser utilizada para aprimorar a triangulação de dados já realizada nesse tipo de estudo, gerando novas conclusões ou respondendo a novas perguntas de pesquisa. Além disso, com base na aplicação da

metodologia em um *corpus* constituído por protocolos verbais dos pesquisadores e dos tradutores, foi possível, utilizando a definição de metarreflexão de Alves (2005, p. 122) – como o "automonitoramento das atividades processuais durante uma tarefa de tradução" –, elaborar uma nova definição com base na Linguística Sistêmico-Funcional. Segundo Halliday e Matthiessen (2014), metarreflexão é a confluência da função SUJEITO, do sistema de SUJEITABILIDADE da metafunção interpessoal, e da função EXPERIENCIADOR, do sistema de TRANSITIVIDADE da metafunção ideacional.

Por fim, caso um pesquisador possua conhecimentos de programação na linguagem *R*, é possível utilizar o *script* como fonte de pesquisa para o desenvolvimento de ferramentas similares ou diferentes, uma vez que essa é uma linguagem de código livre.

SUGESTÕES PARA PESQUISAS FUTURAS

Visto que esses *scripts* possuem o potencial de serem aprimorados para contribuírem de forma mais eficiente para mais pesquisas, sugere-se o aprimoramento dos *scripts* e a análise de dados estruturados e não estruturados visando à análise do processo da tradução por meio da triangulação entre esses dois tipos de dados, o que não foi o objetivo principal desta dissertação. Pode-se também, realizar análises utilizando como base a definição de metarreflexão proposta nesta dissertação em estudos com uma quantidade maior de palavras e/ou mais planilhas de dados a fim de se gerar outros resultados.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, M. A. D. **Estabilidade em análise de agrupamento (Cluster Analysis)**. 2005. 64 f. Dissertação (Mestrado em Biometria) – Departamento de Física e Matemática, Universidade Federal de Pernambuco, Recife, 2005.

ALMEIDA, M. B. Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares. **Scielo**, Brasília, v. 31, n. 2, p. 5-13, maio/ago. 2012.

ALVES, F. (Ed.). **Triangulating Translation:** Perspectives in process oriented research. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2003.

ALVES, F. Ritmo cognitivo, meta-reflexão e experiência: parâmetros de análise processual no desempenho de tradutores novatos e experientes. In: PAGANO, A.; MAGALHÃES, C.; ALVES, F. **Competência em tradução:** cognição e discurso. Belo Horizonte: Editora UFMG, 2005. p. 109-169.

ARANHA, C.; PASSOS, E. A Tecnologia de Mineração de Textos: Artigo Tutorial. **Revista Eletrônica de Sistemas de Informação**, Rio de Janeiro, n. 2, p. 1-8, 2006.

BAKER, M. Towards a Methodology for Investigating the Style of a Literary Translator. **Target**, Amsterdam/Philadelphia, p. 240-266, 2000.

BRAGA, C. N. D. O. **O texto traduzido sob a perspectiva do avaliador:** um estudo exploratório. 2012. 150 f. Dissertação (Mestrado em Linguística Aplicada) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

BRANCO, A. *et al*. Out-of-the-Box Robust Parsing of Portuguese. In: INTERNATIONAL CONFERENCE ON THE COMPUTATIONAL PROCESSING OF PORTUGUESE, 9., 2010, Porto Alegre. **Proceedings of the 9th…** Porto Alegre: [s.n.], 2010. p. 75-85.

BRANCO, A.; SILVA, J. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4., 2004, Lisboa. **Proceedings of the 4th...** Paris: ELRA, 2004. p. 507-510.

CASTRO, R. A. E.; CECÍLIO, S. G. Análise de dados não estruturados - Mineração de textos. In: TORRES, H. D. C.; REIS, I. A.; PAGANO, A. S. **Empoderamento do pesquisador das Ciências da Saúde**. 1a. ed. Belo Horizonte: Tribo da Ilha, 2015. cap. 6, p. 79-97.

CATFORD, J. C. **A linguistic theory of translation:** an essay in applied linguistics. London: Oxford University, 1964.

FEINERER, I.; HORNIK, K.; MEYER, D. Text Mining infrastructure in R. **Journal of Statistical Software**, v. 25, n. 5, p. 1-54, 2008. Disponível em: <http://www.jstatsoft.org/v25i05/paper>.

FERREGUETTI, K. **As orações existenciais em inglês e português brasileiro:** um estudo baseado em *corpus*. 2014. 98 f. Dissertação (Mestrado em Linguística Aplicada) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte, 2014.

FERREGUETTI, K.; RODRIGUES, J. S. N. Transcrição de dados verbais. In: TORRES, H. D. C.; REIS, I. A.; PAGANO, A. S. **Empoderamento do pesquisadores das Ciências da Saúde**. 1a. ed. Belo Horizonte: Tribo da Ilha, 2015. cap. 5, p. 65-78.

FIGUEREDO, G. P. **Introdução ao perfil metafuncional do português brasileiro:** contribuições para estudos multilíngues. 2011. 383 f. Tese (Doutorado em Linguística Aplicada) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

FIGUEREDO, G. P.; PAGANO, A. S.; FERREGUETTI, K. Os sistemas textuais de focalização na organização funcional da gramática do Português Brasileiro. **DELTA**, São Paulo, v. 30, n. 2, p. 309-352, dez. 2014. Disponível em: <http://dx.doi.org/10.1590/0102-44508033430169253 2>.

GRIES, S. T. Useful statistics for corpus linguistics. In: SÁNCHEZ, A.; ALMELA, M. **A mosaic of corpus linguistics:** selected approaches. Frankfurt am Main: Peter Lang, 2010. p. 269-291.

GRIES, S. T. Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us. In: TAAVITSAINEN, I., *et al*. **Developments in English:** expanding electronic evidence. [S.l.]: Cambridge University Press, 2014. p. 29-47.

HAIR, J. F.; BLACK, W. C.; BAB, B. J. **Análise Multivariada de Dados**. Tradução de Adonai Schlup Sant'Anna. 6. ed. Porto Alegre: Bookman, 2009.

HALLIDAY, M. A. K. Grammatical categories in Modern Chinese. **Transactions of the Philological Society**, v. 55, p. 177-224, nov. 1956.

HALLIDAY, M. A. K. **The Language of the Chinese:** Secret History of the Mongols. Oxford: Blackwell, 1959. 235 p. (Publications of the Philological Society, 17).

HALLIDAY, M. A. K. Systemic background. In: BENSON, J. D.; GREAVES, W. S. **Systemic Perspectives on Discourse:** Selected theoretical papers from the 9th International Systemic Workshop. Norwood: Ablex Publishing, v. 1, 1985. p. 1-15.

HALLIDAY, M. A. K. **On Grammar**. Londres; Nova York: Continuum, 2002. (The collected works of M. A. K. Halliday, v. 1).

HALLIDAY, M. A. K. **Computational and quantitative studies**. Londres: Continuum, 2005. (The collected works of M. A. K. Halliday, v. 6).

HALLIDAY, M. A. K.; MATTHIESSEN, C. M. I. M. **Halliday's Introduction to Functional Grammar**. 4a. ed. Oxford: Routledge, 2014.

HALLIDAY, M. A. K.; MCINTOSH, A.; STREVENS, P. **The Linguistic Sciences and Language Teaching**. Londres: Longmans, 1964.

HALLIDAY, M. A. K.; WEBSTER, J. J. Arriving at a theory of the text: A case study of the commencement addresses delivered by Steve Jobs and Susan Rice. In: _____. **Text Linguistics:** The How and Why of Meaning. Sheffield; Bristol: Equinox, 2014. p. 367-425.

HALLIDAY, M. A. K.; WEBSTER, J. J. **Continuum Companion to Systemic Functional Linguistics**. Londres; Nova York: Continuum, 2009.

HAMMARBERG, B. Roles of L1 and L2 in L3 Production and Acquisition. In: CENOZ, J.; HUFEISEN, B.; JESSNER, U. **Cross-Linguistic Influence in Third Language Acquisition:** Psycholinguistic Perspectives. Clevedon: Multilingual Matters, 2001. cap. 2, p. 21-41.

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

JAKOBSEN, A. L. Logging target text production with Translog. In: HANSEN, G. **Probing the process in translation:** methods and results. [S.l.]: Samfundslitteratur, 1999. p. 9-21.

JI, M. Hypothesis testing in corpus-based literary translation studies. In: OAKES, P.; JI, M. **Quantitative Methods in Corpus-Based Translation Studies:** A practical guide to descriptive translation research. [S.l.]: John Benjamins, 2012. v. 51. p. 53-74.

KAZMIER, L. J. **Estatística aplicada à economia e administração**. São Paulo: Pearson Makron, 2004.

KERNS, G. J. **Introduction to Probability and Statistics Using R**. 1a. ed. [S.l.]: [s.n.], 2011. Disponível em: <https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf>. Acesso em: 24 maio 2016.

KNIGHT, N.; MAHBOOK, A. (Eds.). **Appliable Linguistics**. Sydney: Continuum, 2010. p. 1-12.

LANE, D. M. Estimation. In: LANE, D. M., *et al.* **Online Statistics Education:** A Multimedia Course of Study. [S.l.]: Rice University. cap. 10, p. 334. Disponível em: <http://onlinestatbook.com/>. Acesso em: 24 maio 2016.

LIMA, K. C. S. **Caracterização de registros orientada para a produção textual no ambiente multilíngue:** Um estudo baseado em corpora comparáveis. 2013. 251 f. Tese (Doutorado em Linguística Aplicada) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

LINDEN, R. Técnicas de Agrupamento. **Revista de Sistemas de Informação da FSMA**, Santa Maria, v. 4, p. 18-36, 2009. Disponível em: <http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf>. Acesso em: 4 fev. 2016.

LOURENÇO, I. A. **Conhecimento experto em tradução:** aferição da durabilidade de tarefas tradutórias realizadas por sujeitos não-tradutores em condições empírico-experimentais. 2007. 277 f. Dissertação (Mestrado em Linguística Aplicada) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

LOURENÇO, I. A. **(Des)compactação de significados e esforço cognitivo no processo tradutório:** um estudo da metáfora gramatical na construção do texto traduzido. 2012. 295 f. Tese (Doutorado em Linguística Aplicada) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

MAGALHÃES, C. M.; ALVES, F. Investigando o Papel do Monitoramento Cognitivo-Discursivo. **Cadernos de Tradução**, Florianópolis, v. 17, p. 71-127, 2006.

MALMKJAER, K. Translational stylistics: Dulcken's translations of Hans Christian Andersen. **Language and Literature**, v. 13, n. 13, p. 13-24, 2004. Disponível em: <http://lal.sagepub.com/cgi/content/abstract/13/1/13>.

MALMKJAER, K. **Linguistics and the Language of Translation**. Edinburgh: Edinburgh University Press, 2005.

MALTA, G. **O processamento cognitivo em tarefas de re(tradução):** um estudo baseado em rastreamento ocular, registro de teclado e mouse e protocolos retrospectivos. 2015. 251 f. Tese (Doutorado em Linguística Aplicada) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

MATTHIESSEN, C. M. I. M. Systemic Functional Linguistics as appliable linguistics: social accountability and critical approaches. **DELTA**, São Paulo, v. 28, p. 435-471, 2012.

MATTHIESSEN, C. M. I. M.; TERUYA, K.; LAM, M. **Key Terms in Systemic Functional Linguistics**. Londres: Continuum, 2010.

NUNES, L. P. **Relações coesivas e estruturais:** um estudo de conjunções em córpus paralelo e comparável no par linguístico inglês–português brasileiro. 2014. 273 f. Tese (Doutorado em Linguística Aplicada) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte, 2014.

OAKES, M.; JI, M. (Eds.). **Quantitative Methods in Corpus-Based Translation Studies:** A practical guide to descriptive translation research. Amsterdã/Filadélfia: John Benjamins, 2012.

OAKES, P. Describing a translational corpus. In: OAKES, P.; JI, M. **Quantitative Methods in Corpus-Based Translation Studies:** A practical guide to descriptive translation research. [S.l.]: John Benjamins, 2012. v. 51. p. 115-148.

PAGANO, A.; FIGUEREDO, G. P.; LUKIN, A. **Empirical Translation Studies:** Interdisciplinary Methodologies Explored. [S.l.]: Equinox Publishing Ltd., 2014.

PAGANO, A.; VASCONCELLOS, M. L. Explorando interfaces: estudos da tradução, linguística sistêmico funcional e linguística de córpus. In: ALVES, F.; MAGALHÃES, C.; PAGANO, A. **Competência em tradução:** cognição e discurso. Belo Horizonte: Editora UFMG, 2005. p. 177-207.

R: A language and environment for statistical computing. Versão 3.3.0. Vienna: R Foundation for Statistical Computing, 2016. Disponível em: <http://www.R-project.org/>.

REIS, I. A. Dados, informação, conhecimento nas Ciências da Saúde: do não estruturado ao estruturado. In: TORRES, H. D. C.; REIS, I. A.; PAGANO, A. S. (Orgs.). **Empoderamento do pesquisador das ciências da saúde**. 1a. ed. Belo Horizonte: Tribo da Ilha, 2015. cap. 3, p. 37-47.

REZENDE, S. O. (Ed.). **Sistemas Inteligentes:** Fundamentos e Aplicações. Barueri: Editora Manole, 2003.

RYBICKI, J. The great mystery of the (almost) invisible translator: Stylometry in translation. In: OAKES, P.; JI, M. **Quantitative Methods in Corpus-Based Translation Studies:** A practical guide to descriptive translation research. [S.l.]: John Benjamins, 2012. v. 51. p. 231-248.

SARDINHA, T. B. Lingüística de Corpus: histórico e problemática. **DELTA**, São Paulo, v. 16, n. 2, 2000. 323-367.

SCHEUREN, F. What is Margin of Error. In: SCHEUREN, F. **What is a survey**. [S.l.]: [s.n.], 2004. cap. 10. Disponível em: <http://www.amstat.org/sections/srms/pamphlet.pdf>. Acesso em: 24 maio 2016.

SILVA, J. M. G. **Relações de equivalência português brasileiro-inglês:** um estudo de caso. Dissertação (Mestrado em Estudos Linguísticos) – Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte. Em andamento.

STUBBLEBINE, T. **Regular Reference Pocket Reference**. 2a. ed. Sebastopol: O'Reilly, 2007.

SWATHI, P.; YOGISH, H. K.; SREERAJ, R. S. Predictive Data Mining Procedures for the Prediction of Coronary Artery Disease. **International Journal of Emerging Technology and Advanced Engineering**, v. 5, n. 2, fev. 2015. Disponível em: <http://www.ijetae.com/files/Volume5Issue2/IJETAE_0215_61.pdf>. Acesso em: 12 jan. 2016.

TOURY, G. **In Search of Theory of Translation**. Tel Aviv: Porter Institute for Poetics and Translation, 1980.

VENUTI, L. Translation, Community, Utopia. In: VENUTI, L.; BAKER, M. **The Translation Studies Reader**. Londres/Nova York: Routledge, 2000. cap. 30, p. 468-488.

WHITNEY, H. How to define data, information and knowledge. **TechTarget**, 2007. Disponível em: <http://searchdatamanagement.techtarget.com/feature/Defining-data-information-and-knowledge>. Acesso em: 25 fev. 2016.

ZARA, J. V. Divulgação do R para linguistas. In: ENCONTRO VIRTUAL DE DOCUMENTAÇÃO EM SOFTWARE LIVRE E CONGRESSO INTERNACIONAL DE LINGUAGEM E TECNOLOGIA ONLINE, 10., 2013, Belo Horizonte. **Anais do...** Belo Horizonte: Faculdade de Letras/UFMG, 2013.

ANEXOS

ANEXO 1 – Questionário prospectivo do pesquisador R2

CD 34897

1

QUESTIONÁRIO PROSPECTIVO

Projeto Expert@ – Conhecimento experto em tradução: modelagem do processo tradutório em altos níveis de desempenho (CNPq 479340/2006-0).

Pesquisador responsável: Prof. Dr Fabio Alves (UFMG)
Projeto aprovado e registrado no SISNEP (Sistema Nacional de Ética em Pesquisa) sob o número CAAE - 0033.0.203.000-05

Sigla: R2 (espaço a ser preenchido pelo pesquisador)
Grupo: Revisão (espaço a ser preenchido pelo pesquisador)
Data: 18/12/2008

Visão
1. Visão corrigida (uso de lentes, cirurgia etc.)?
(X) Sim. Causa: _____ (se tiver feito cirurgia, passe para a pergunta 1).
( ) Não (passe para a pergunta 4).

2. Está usando alguma lente corretiva agora?
(X) Sim.
( ) Não (passe para a pergunta 4).

3. Está utilizando:
(X) óculos.
( ) lentes.

Perfil lingüístico

4. Qual o seu nível de conhecimento de línguas estrangeiras?

4A. Idioma – Inglês

[ ] Lê	( ) Pouco	( ) Razoavelmente	(X) Bem
[ ] Fala	( ) Pouco	( ) Razoavelmente	(X) Bem
[ ] Escreve	( ) Pouco	( ) Razoavelmente	(X) Bem
[ ] Compreende	( ) Pouco	( ) Razoavelmente	(X) Bem

4B. Outro Idioma (especificar: Espanhol)

[ ] Lê	( ) Pouco	( ) Razoavelmente	(X) Bem
[ ] Fala	( ) Pouco	( ) Razoavelmente	(X) Bem
[ ] Escreve	( ) Pouco	( ) Razoavelmente	(X) Bem
[ ] Compreende	( ) Pouco	( ) Razoavelmente	(X) Bem

4C. Outro Idioma (especificar: Alemão)

[ ] Lê	(X) Pouco	( ) Razoavelmente	( ) Bem
[ ] Fala	(X) Pouco	( ) Razoavelmente	( ) Bem
[ ] Escreve	(X) Pouco	( ) Razoavelmente	( ) Bem
[ ] Compreende	(X) Pouco	( ) Razoavelmente	( ) Bem

5. Já residiu em algum país de fala inglesa?
( ) Não.
(X) Sim. Especificar por quanto tempo: EUA/Inglaterra 4 meses 4 meses

6. Já residiu em algum outro país?
(X) Não.
( ) Sim. Especificar qual e por quanto tempo: _____.

Hábitos de leitura e redação

7. Você tem hábito de leitura regular em português?
( ) Não.
(X) Sim. Especificar tipos de texto: Romance, jornal, revista (Carta Capital)

8. Você tem hábito de leitura regular em inglês?
( ) Não.
(X) Sim. Especificar tipos de texto: Ficção, revista e livro

9. Você tem hábito de leitura regular em alguma outra língua estrangeira além do inglês?
( ) Não.
(X) Sim. Especificar tipos de texto: Romance, Jornal (Espanhol).

10. Você costuma escrever com regularidade em português?
( ) Não.
(X) Sim. Especificar tipos de texto: E-mail, relatório técnico

11. Você costuma escrever com regularidade em inglês?
( ) Não.
(X) Sim. Especificar tipos de texto: artigo e e-mails.

12. Você costuma escrever com regularidade em algumu outru língua estrangeira além do inglês?
( ) Não.
(X) Sim. Especificar tipos de texto: E-mail (Espanhol).

13. Como você aprendeu a redigir textos acadêmicos em português?
( ) Fazendo algum curso.
( ) Copiando de outros textos.
( ) Com ajuda de algum colega.
( ) Com ajuda do orientador de dissertação e/ou tese.
(X) Outros. Especificar: Comparando com outro texto

14. Como você aprendeu a redigir textos acadêmicos em inglês?
( ) Fazendo algum curso.
( ) Copiando de outros textos.
( ) Com ajuda de algum colega.
( ) Com ajuda do orientador de dissertação e/ou tese.
(X) Outros. Especificar: Comparando.

Divulgação científica

15. Você lê revistas ou livros de divulgação científica em português?
( ) Nunca.
( ) Raramente.
( ) Às vezes.
(X) Freqüentemente. Fapemig
( ) Muito freqüentemente.
( ) Sempre.

16. Você lê revistas ou livros de divulgação científica em inglês?
( ) Nunca.
(X) Raramente.
( ) Às vezes.
( ) Freqüentemente.
( ) Muito freqüentemente.
( ) Sempre.

17. Você lê revistas ou livros de divulgação científica em alguma outra língua estrangeira além do inglês?
(X) Nunca (passe para a pergunta 19).
( ) Raramente.
( ) Às vezes.
( ) Freqüentemente.
( ) Muito freqüentemente.
( ) Sempre.

18. Especifique a(s) língua(s), além do inglês, em que é(são) veiculada(s) a(s) divulgação(ões) científica(s) que você lê.

_____.

13. Como você aprendeu a redigir textos acadêmicos em português?
( ) Fazendo algum curso.
( ) Copiando de outros textos.
( ) Com ajuda de algum colega.
( ) Com ajuda do orientador de dissertação e/ou tese.
(X) Outros. Especificar: Comparando com outro texto

14. Como você aprendeu a redigir textos acadêmicos em inglês?
( ) Fazendo algum curso.
( ) Copiando de outros textos.
( ) Com ajuda de algum colega.
( ) Com ajuda do orientador de dissertação e/ou tese.
(X) Outros. Especificar: Comparando.

Divulgação científica

15. Você lê revistas ou livros de divulgação científica em português?
( ) Nunca.
( ) Raramente.
( ) Às vezes.
(X) Freqüentemente.
( ) Muito freqüentemente.
( ) Sempre.

Papermag

16. Você lê revistas ou livros de divulgação científica em inglês?
( ) Nunca.
(X) Raramente.
( ) Às vezes.
( ) Freqüentemente.
( ) Muito freqüentemente.
( ) Sempre.

17. Você lê revistas ou livros de divulgação científica em alguma outra língua estrangeira além do inglês?
(X) Nunca (passe para a pergunta 19).
( ) Raramente.
( ) Às vezes.
( ) Freqüentemente.
( ) Muito freqüentemente.
( ) Sempre.

18. Especifique a(s) língua(s), além do inglês, em que é(são) veiculada(s) a(s) divulgação(ões) científica(s) que você lê.

_____.

19. Indique alguns títulos de revistas ou livros de divulgação científica que você tenha lido, em português e/ou língua estrangeira, nos últimos dois anos.
Mundo Estranho, Ciência, Nat Geo

20. Você tem hábito de assistir a documentários científicos?
(X) Nunca.
( ) Raramente.
( ) Às vezes.
( ) Freqüentemente.
( ) Muito freqüentemente.
( ) Sempre.

21. Você já fez parte de algum projeto de divulgação científica?
(X) Sim.
( ) Não (passe para a pergunta 23).

22. Especifique o(s) tipo(s) de divulgação de que você já participou.
(X) Reportagens de jornal
( ) Reportagens na TV
(X) Reportagens no rádio.
( ) Textos para divulgação na Internet.
(X) Outro. Especificar: COM pares alunos

Avaliação de textos científicos

23. Quais são os principais atributos de um bom texto científico? Numere os itens em cada aspecto descrito a seguir do mais importante (1) ao menos importante.
(4) Termos técnicos específicos.
(5) Correção gramatical.
(3) Informações suficientes.
(1) Objetivos claros.
(2) Metodologia clara e consistente.
(6) Outro. Especificar: Conclusão .

24. Em um texto científico bem redigido em português, o(s) autor(es):
(X) não aparecem de forma explícita (através, por exemplo, de expressões como "nós" ou de verbos conjugados na primeira pessoa)
( ) podem aparecer de forma explícita em algumas seções ou em alguns capítulos do texto (através, por exemplo, de expressões como "nós" ou de verbos conjugados na primeira pessoa)
( ) devem aparecer de forma explícita dependendo do tipo de texto (através, por exemplo, de expressões como "nós" ou de verbos conjugados na primeira pessoa)
( ) sempre aparecem de forma explícita (através, por exemplo, de expressões como "nós" ou de verbos conjugados na primeira pessoa)

19. Indique alguns títulos de revistas ou livros de divulgação científica que você tenha lido, em português e/ou língua estrangeira, nos últimos dois anos.
Mundo Estranho, Ciência, Nat Geo

20. Você tem hábito de assistir a documentários científicos?
(X) Nunca.
( ) Raramente.
( ) Às vezes.
( ) Freqüentemente.
( ) Muito freqüentemente.
( ) Sempre.

21. Você já fez parte de algum projeto de divulgação científica?
(X) Sim.
( ) Não (passe para a pergunta 23).

22. Especifique o(s) tipo(s) de divulgação de que você já participou.
(X) Reportagens de jornal
( ) Reportagens na TV
(X) Reportagens no rádio.
( ) Textos para divulgação na Internet.
(X) Outro. Especificar: COM pares alunos

Avaliação de textos científicos

23. Quais são os principais atributos de um bom texto científico? Numere os itens em cada aspecto descrito a seguir do mais importante (1) ao menos importante.
(4) Termos técnicos específicos.
(5) Correção gramatical.
(3) Informações suficientes.
(1) Objetivos claros.
(2) Metodologia clara e consistente.
(6) Outro. Especificar: Conclusão .

24. Em um texto científico bem redigido em português, o(s) autor(es):
(X) não aparecem de forma explícita (através, por exemplo, de expressões como "nós" ou de verbos conjugados na primeira pessoa)
( ) podem aparecer de forma explícita em algumas seções ou em alguns capítulos do texto (através, por exemplo, de expressões como "nós" ou de verbos conjugados na primeira pessoa)
( ) devem aparecer de forma explícita dependendo do tipo de texto (através, por exemplo, de expressões como "nós" ou de verbos conjugados na primeira pessoa)
( ) sempre aparecem de forma explícita (através, por exemplo, de expressões como "nós" ou de verbos conjugados na primeira pessoa)

25. Em um texto científico bem redigido em inglês ou traduzido para o inglês, o(s) autor(es):
(X) não aparecem de forma explícita (através, por exemplo, de expressões como "we" ou "us")
( ) podem aparecer de forma explícita em algumas seções ou em alguns capítulos do texto (através, por exemplo, de expressões como "we" ou "us").
( ) devem aparecer de forma explícita dependendo do tipo de texto (através, por exemplo, de expressões como "we" ou "us").
( ) sempre aparecem de forma explícita (através, por exemplo, de expressões como "we" ou "us").

Tradução/Redação/Revisão

26. Nos últimos dois anos, quantos artigos ou resumos em português você traduziu/redigiu ou ajudou a traduzir/redigir para o inglês.
( ) Nenhum.
( ) Até 2.
(X) De 2 a 5.
( ) Mais de 5.

27. Qual a porcentagem desses textos foram publicados ou já receberam aceite?
( ) Até 40%
(X) De 40% a 70%
( ) Acima de 70%

28. Quando tem dúvidas durante a tradução para o inglês ou durante a redação em inglês, a quais fontes você recorre com mais freqüência?
( ) Dicionário bilíngüe.
( ) Dicionário monolíngüe.
( ) Dicionário técnico.
( ) Ajuda de colega(s).
(X) Outra(s) fonte(s) de referência. Especificar: Internet, artigos em inglês, glossário em inglês

ANEXO 2 - Questionário prospectivo do tradutor T2

Questionário para Entrevista Prospectiva

Projeto Expert@ – Conhecimento experto em tradução: modelagem do processo tradutório em altos níveis de desempenho (CNPq 479340/2006-4).

Pesquisador responsável: Prof Dr Fabio Alves (UTMG)
Projeto aprovado e registrado no SISNEP (Sistema Nacional de Ética em Pesquisa) sob o número CAAE - 0033.0.203.000-05

Questionário/Entrevista Prospectiva

QUESTIONÁRIO AOS TRADUTORES DE LÍNGUA INGLESA (adaptado de Durão, 2005)

Este questionário visa conhecer o perfil dos tradutores informantes da pesquisa em andamento.
Sua contribuição é muito importante.
Atenciosamente,
Grupo EXPERT@

[redacted]

Visão

A. Visão corrigida?
(X) Sim. Causa: Astigmatismo
( ) Não (passe para a pergunta 1)

B. Está usando alguma lente corretiva agora?
(X) Sim.
( ) Não (passe para a pergunta 1).

C. Está utilizando:
(X) Óculos
( ) Lentes

Perfil acadêmico

1- Tem bacharelado, licenciatura ou grau equivalente?
(X) Sim
( ) Não (passe para a pergunta 13)

2- Em que área(s)?
Favor especificar: Engenharia Química

3- Formou-se:
(X) no Brasil?
( ) Outro(s) país(es)? (por favor especifique) _____

4- Fez alguma pós-graduação *lato-sensu* (especialização)?
(X) Sim Está fazendo
( ) Não (passe para a pergunta 7)

5. Fez pós-graduação/ções *lato-sensu* em:
(X) Tradução
( ) Outra(s) área(s): (por favor especifique) _____

6- Fez a(s) sua(s) pós-graduação/ções *lato-sensu*:
(X) no Brasil?
( ) Outro(s) país(es)? (por favor especifique) _____

7- Tem mestrado?
( ) Sim
(X) Não (passe para a pergunta 10)

8- É mestre em:
( ) Tradução
( ) Outra(s) área(s): (por favor especifique) _____

9- Fez o seu mestrado:
( ) no Brasil?
( ) Outro(s) país(es)? (por favor especifique) _____

10- Tem doutoramento?
( ) Sim
(X) Não (passe para a pergunta 13)

11- É doutor em:
( ) Tradução
( ) Outra(s) área(s): (por favor especifique) _____

12- Fez o seu doutoramento:
( ) no Brasil?
( ) Outro(s) país(es)? (por favor especifique) _____

Perfil lingüístico

13- Qual é a sua língua materna?
(X) Português
( ) Bilíngüe (Português e outra)
( ) Outra que não o português (se marcou esta opção, pare aqui)

14- Especifique o nível de conhecimento do idioma de trabalho (inglês):
(X) bilíngüe
( ) muito proficiente
( ) proficiente
( ) pouco proficiente

15-Já residiu em país em que a sua língua de trabalho é falada predominantemente?
(X) Não
( ) Sim (especifique por quanto tempo: _____)

16- Tem conhecimento de outros idiomas além da língua de trabalho?
(X) Sim (por favor especifique o(s) idioma(s) Alemão
( ) Não

Perfil profissional

17- Há quantos anos é tradutor?
(X) até 2 anos
( ) de 2 a 4 anos
( ) de 4 a 6 anos
( ) de 6 a 10 anos
( ) mais de 10 anos

18- É tradutor:
(X) independente
( ) em escritório de tradução
( ) ambos

19- A tradução é a sua atividade principal?
(X) Sim (passe para a pergunta 21)
( ) Não

20- Qual a sua principal atividade profissional?
Favor especificar: _____

21- Que percentagem do seu rendimento provém da tradução?
( ) até 40%
( ) de 40% a 70%
(X) acima de 70%

22- Que percentagem de suas traduções é feita no par lingüístico inglês>português?
( ) até 40%
(X) de 40% a 70%
( ) acima de 70%

23- Volume de material traduzido no par lingüístico português > inglês nos últimos dois (2) anos? (favor indicar uma média em número de laudas traduzidas por dia)

Aprox. _____ laudas[1]/dia. 3-4.000 palavras

24-Volume de material traduzido no par lingüístico inglês > português nos últimos dois (2) anos? (favor indicar uma média em número de laudas traduzidas por dia)
Aprox. _____ laudas[1]/dia. 4-5.000 palavras

25 - Que tipo de texto você traduz mais freqüentemente?
(X) Técnico
( ) Científico
( ) Literário
( ) Outros (por favor especifique) _____

26A- De que outros idiomas você traduz?
( ) Alemão
( ) Espanhol
( ) Francês
( ) Outro (por favor especifique) _____

26A- Para que outros idiomas você traduz?
( ) Alemão
( ) Espanhol
( ) Francês
( ) Outro (por favor especifique) _____

Aperfeiçoamento profissional

27- Nos últimos dois anos, freqüentou cursos de formação ou atualização profissional em tradução?
(X) Sim
( ) Não (passe para a pergunta 31)

28- Quantos desses cursos freqüentou:
( ) até 2
(X) de 2 a 5
( ) mais de 5

29- Em sua maioria, estes cursos ocorreram:
(X) no Brasil
( ) Outros países

30- Esses cursos foram promovidos por:
( ) Agências de tradução
( ) Empresas
( ) Escolas de idiomas
( ) Instituições de ensino superior público
( ) Instituições de ensino superior privado
( ) Outras entidades
Tradutores

[1] 1 lauda = 1625 caracteres (incluindo espaços).

Sistemas de Memória de Tradução (SMT)

31- Trabalha com SMTs?
(X) Sim
( ) Não

32- Há quanto tempo?
(X) até 2 anos
( ) de 2 a 4 anos
( ) de 4 a 6 anos
( ) de 6 a 10 anos
( ) mais de 10 anos

33- Qual SMT você utiliza?
( ) Trados
( ) Outros (por favor especifique) __Wordfast__

34- Já utilizou outros sistemas?
(X) Sim (por favor especifique) __Trados (poucas vezes)__
( ) Não

Material e condições de trabalho

35- Quais fontes de documentação utiliza com mais freqüência?
( ) dicionário bilíngüe
(X) dicionário monolíngue
(X) dicionário técnico
( ) outras fontes de referência (por favor especifique): _____

36- Recorre a revisores profissionais para tradução do português para o inglês?
(X) Não
( ) Sim

37- Recorre a revisores profissionais para a tradução do inglês para o português?
(X) Não
( ) Sim

Conhecimentos sobre Tradução

38 - O que você considera deva ser priorizado ao se traduzir um texto?
Numere os itens em cada aspecto descrito a seguir em ordem hierárquica crescente:
Aspecto 1
(2) Resolver problemas de linguagem
(1) Resolver problemas relacionados ao conteúdo do texto

Aspecto 2
(1) Dar atenção aos significados do texto de partida
→ (2) Dar atenção à recepção do texto traduzido pelo público-alvo

Aspecto 3
(2) Resolver dúvidas com base nos seus próprios conhecimentos (linguísticos e extra-lingüísticos)
(1) Resolver dúvidas através de buscas em fontes externas (Internet, dicionários impressos).