Rev. Estud. Ling., Belo Horizonte, v. 28, n. 1, p. 13-40, 2020

Elementos de coesão no *Corpus de Língua Portuguesa em Tradução*: investigando a classe gramatical *conjunção* numa perspectiva contrastiva linguística e textual

Cohesive devices in the Corpus de Língua Portuguesa em Tradução*: investigating conjunctions in a contrastive perspective within a text typology*

Leonardo Pereira Nunes

Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais / Brasil

leopereiranunes@gmail.com

https://orcid.org/0000-0002-0678-7137

Resumo: Esta contribuição perfaz análise de elementos de coesão em textos originais e traduzidos, averiguando o impacto das tipologias textual e linguística na frequência de ocorrência de itens da classe gramatical *conjunção* (NUNES, 2014). Dados foram obtidos a partir do *Corpus de Língua Portuguesa em Tradução*, um corpus paralelo bilíngue bidirecional no par linguístico inglês-português brasileiro composto por oito tipos de texto: artigo acadêmico, discurso político, divulgação científica, ficção, manual de instrução, propaganda turística, resenha e website educacional. Utilizou-se o *TreeTagger* para anotação morfossintática e o ambiente de programação R para extração automática e tratamento estatístico das frequências. Verificaram-se frequências significativamente acima das esperadas em textos dos tipos *resenha* e *discurso político*, corroborando a hipótese sobre a explicitação de marcas conjuntivas em textos argumentativos. Ainda, os achados parcialmente confirmaram a hipótese da explicitação significativamente acima da esperada nos textos traduzidos e nos textos originais e traduzidos em português brasileiro. Também revelaram significâncias estatísticas proeminentes nas frequências obtidas em textos dos tipos *ficção* e *website educacional*, apontando nestes tendência à explicitação de conjunções nos textos traduzidos em inglês e naqueles a mesma tendência nos textos originais e traduzidos em português brasileiro. Os resultados dessa investigação sobretudo contribuem para os estudos descritivos da tradução no que tange à descrição linguística do inglês e do português brasileiro em seus modos escritos.

Palavras-chave: *Corpus de Língua Portuguesa em Tradução*; tipologia linguística; tipologia textual; conjunção.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.28.1.13-40

Abstract: This contribution delves into the investigation of cohesive devices in original and translated texts by querying the impact of text and language typologies on the frequency of conjunctions (NUNES, 2014). Data was obtained in the *Corpus de Língua Portuguesa em Tradução,* a bilingual bidirectional parallel corpus in the language pair English-Brazilian Portuguese. The corpus is comprised of eight text types: research article, political speech, science popularisation, fiction, instruction manual, tourism leaflet, review and educational website. *TreeTagger* was used for POS tagging and R environment utilized to perform automatic word frequency and significance testing. The results showed highly above expected frequencies in *reviews* and in *political speeches*, thus corroborating explicitation hypotheses as to the frequency of cohesive marks in argumentative texts. Also, the explicitation hypothesis as to significantly above expected frequencies of conjunctions in translated texts and in the original and translated texts in Brazilian Portuguese was partially corroborated. Findings also showed relevance in statistically significant frequencies in fictional and educational website texts. As to the former, a tendency for the explicitation of conjunctions in original and translated texts in Brazilian Portuguese was revealed. Conversely, frequencies in the latter pointed to a tendency for the explicitation of conjunctive marks in translated texts in English. The findings mostly contribute to descriptive translation studies concerning language description of English and Brazilian Portuguese in their written modes.

Keywords: *Corpus de Língua Portuguesa em Tradução*; language typology; text typology; conjunction.

Recebido em 26 de março de 2019

Aceito em 17 de junho de 2019

1. Introdução

Este trabalho reporta uma investigação conduzida por Nunes (2014), e apresenta uma análise automática de elementos de coesão numa perspectiva interlinguística e numa tipologia de textos. Pontualmente, percorre o escrutínio da frequência de ocorrência de itens da classe de palavra *conjunção* realizado no *Corpus de Língua Portuguesa em Tradução*, um corpus paralelo bilíngue nas direções inglês-português brasileiro e português brasileiro-inglês.

O estudo se insere no âmbito dos estudos da tradução puros descritivos orientados ao produto (cf. HOLMES, 1972), uma vez que fornece resultados de investigação de sistemas linguísticos em contato na relação tradutória e em comparação entre distintos tipos de texto,

escrutinizando características (no que toca a frequência) de elementos coesivos estabelecidos por conjunções inseridas numa tipologia de textos.

Para tal, se vale de uma abordagem inovadora sobretudo no que concerne recursos metodológicos utilizados para o levantamento, processamento e tratamento de dados com fundamento estatístico.

Este artigo está organizado em 5 seções, além desta Introdução. A segunda seção discorre sobre um breve arcabouço teórico que percorre elementos de coesão nas línguas inglesa e portuguesa. A terceira seção apresenta e descreve o corpus sob escrutínio. A quarta seção explicita os procedimentos de anotação e análise automatizada do corpus, bem como a abrangência da referida investigação. A quinta seção reporta os resultados e discussões, e, por fim, a sexta seção tece as conclusões da pesquisa.

2. Elementos de coesão nas línguas inglesa e portuguesa

Halliday e Hasan (1976) advogam que a coesão no inglês é estabelecida por quatro categorias, quais sejam: organização (ou coesão) lexical, referência, substituição/elipse e conjunção. A coesão lexical tem o léxico como elemento chave e se estabelece "através da escolha de itens que se relacionam em um texto através de palavras isoladas ou unidades maiores, como o grupo nominal", por exemplo (HALLIDAY; MATTHIESSEN, 2014, p. 606 *apud* NUNES, 2014). A referência, por sua vez, "estabelece na esfera gramatical uma cadeia de elementos que se relacionam intra e extratextualmente através de itens endofóricos e exofóricos, respectivamente" (HALLIDAY; MATTHIESSEN, 2014, p. 606 *apud* NUNES, 2014). A substituição e a elipse também possuem caráter gramatical e compreendem "ferramentas que permitem a exclusão de partes de uma estrutura se estas puderem ser inferidas através de elementos antecedentes no texto" (HALLIDAY; MATTHIESSEN, 2014, p. 606 *apud* NUNES, 2014). Por fim, conjunções, foco deste estudo, são apreciadas enquanto "instrumentos sistemáticos de conexão de orações e complexos oracionais" (HALLIDAY; MATTHIESSEN, 2014, p. 609 *apud* NUNES, 2014).

Na variante brasileira da língua portuguesa, Neves (2011) discorre sobre a coesão textual sobremaneira ao percorrer elementos sobre a referência (situacional e textual) e sobre as conjunções coordenativas e subordinativas adverbiais, estas vislumbradas enquanto instrumentos sequenciadores e de amarração de blocos de textos (NEVES, 2011, p. 19).

Como explicitado anteriormente, apenas elementos da classe de palavra *conjunção* foram automaticamente investigados no corpus da pesquisa, apresentado na próxima seção.

3. O *Corpus de Língua Portuguesa em Tradução*

O *Corpus de Língua Portuguesa em Tradução* (doravante Klapt!)[1] foi compilado a partir do corpus CroCo[2] (*Cross-linguistic corpora*) (cf. NEUMANN, 2005, 2008) por pesquisadores do Laboratório Experimental de Tradução (LETRA)[3] da Universidade Federal de Minas Gerais (UFMG). O intuito foi o de se investigar elementos linguísticos numa perspectiva contrastiva e considerndo-se múltilos tipos de textos, no par linguístico inglês-português brasileiro e em ambas as direções.

O Klapt! pode ser categorizado como um corpus multilíngue paralelo bidirecional (GRANGER, 2003, p. 21), sendo possíveis análises linguísticas tanto entre os textos originais e suas respectivas traduções nas direções português brasileiro-inglês e inglês-português brasileiro. Dado o desenho bidirecional do corpus, pode-se, ainda, investigá-lo nas perspectivas comparáveis mono e bilíngues de textos originais e de textos traduzidos.

A tipologia textual contemplada para a criação do corpus baseou-se em processos sociossemióticos que, em síntese, resumem a maneira como a língua, e todo o seu potencial de criação de significados, são instanciados no contexto de cultura (cf. MATTHIESSEN; TERUYA; LAM, 2010; HALLIDAY; MATTHIESSEN, 2014). Esses processos incluíram diversos tipos de texto da língua escrita definidos por distintas

---

[1] Este acrônico foi cunhado a partir de um intertexto fonológico com verbo klappt na língua alemã, que corresponde ao verbo funcionar na língua portuguesa (NUNES, 2010)

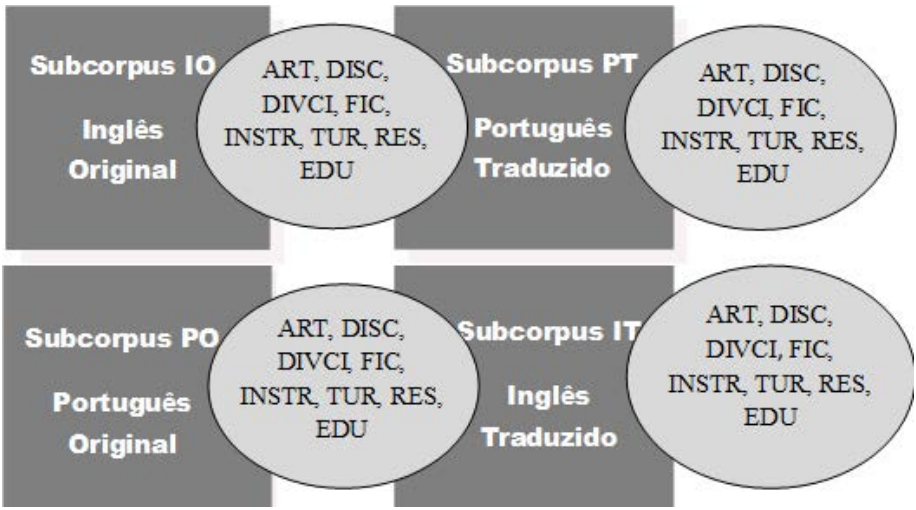
[2] Esse projeto, desenvolvido por pesquisadores da Universidade do Sarre (Alemanha), teve por objetivo identificar as especificidades do texto traduzido em comparação ao texto não traduzido (incluindo a explicitação e outras propriedades da tradução) entre o inglês e o alemão, e em ambas as direções. Página do projeto: http://fedora.clarin-d.uni-saarland.de/croco-gecco/croco/presentation_neumann_hansenschirra.pdf. Acesso em 12 mar. 2019.

[3] O corpus está disponibilizado no Portal Min@s, um ambiente virtual que agrega bancos de dados de laboratórios vinculados ao Programa de Pós-Graduação em Estudos Linguísticos (POSLIN) da Faculdade de Letras (FALE) da UFMG. Endereço do portal: http://portalminas.letras.ufmg.br/. Acesso em 12 mar. 2019.

comunidades de usuários, sendo contemplados oito tipos: artigos acadêmicos, discursos políticos, textos de divulgação científica, textos de ficção, manuais de instrução, textos de propagandas turísticas, resenhas e textos de *websites* educacionais.

A Figura 1 apresenta um esquema gráfico do corpus.

FIGURA 1 – Desenho do Klapt!



Fonte: Nunes (2014, p. 73)

Quanto ao tamanho, cada tipo textual em cada um dos 4 subcorpora possui uma média de 10 textos com 3.000 palavras (*tokens*)[4] correntes cada, totalizando amostras de aproximadamente 30.000 palavras, conforme ilustrado na TABELA 1.

[4] Cumpre mencionar que as amostras selecionadas compreenderam textos na íntegra ou excertos contendo parágrafos inteiros, de modo que a coesão textual pudesse ser mantida.

TABELA 1 – Números de *tokens* do Klapt! por tipo textual e subcorpus

Tipo textual	Inglês original (IO)	Inglês traduzido (IT)	Português original (PO)	Português traduzido (PT)	Total por tipo textual
Artigo acadêmico	30.299	30.163	30.049	31.629	**122.140**
Discurso político	30.178	30.587	29.813	31.080	**121.658**
Divulgação científica	30.664	32.749	30.790	31.010	**125.213**
Ficção	30.138	32.955	30.072	30.881	**124.046**
Manual de instrução	29.453	28.527	29.244	35.628	**122.852**
Propaganda turística	27.871	30.474	30.191	28.487	**117.023**
Resenha	30.126	31.959	32.052	30.960	**125.097**
Website educacional	29.828	28.131	29.100	32.322	**119.381**
Total por subcorpus	**238.557**	**245.54**	**241.311**	**251.997**	
Total geral	**977.410**				

Fonte: Nunes (2014, p. 75)

Como mostra a Tabela 1, o Klapt! como um todo totaliza aproximadamente 980 mil palavras, o que o caracteriza como um corpus de extensão média, isto é, entre 250 mil e 1 milhão de palavras (BERBER SARDINHA, 2004, p. 26).

No que tange às suas aplicações, o Klapt! pode ser utilizado enquanto recurso para diversas pesquisas no âmbito dos estudos da tradução (tanto puros quanto aplicados), tais como em investigações das propriedades da tradução, em pesquisas orientadas ao processo e ao produto tradutórios, no desenvolvimento de metodologias de anotação multidimensional, na análise de registro, na descrição linguística e na formação de tradutores (cf. JESUS; NUNES, 2014).

A próxima seção discorre sobre os procedimentos de anotação e extração automática com subsídio estatístico da frequência de ocorrência de conjunções no corpus Klapt!

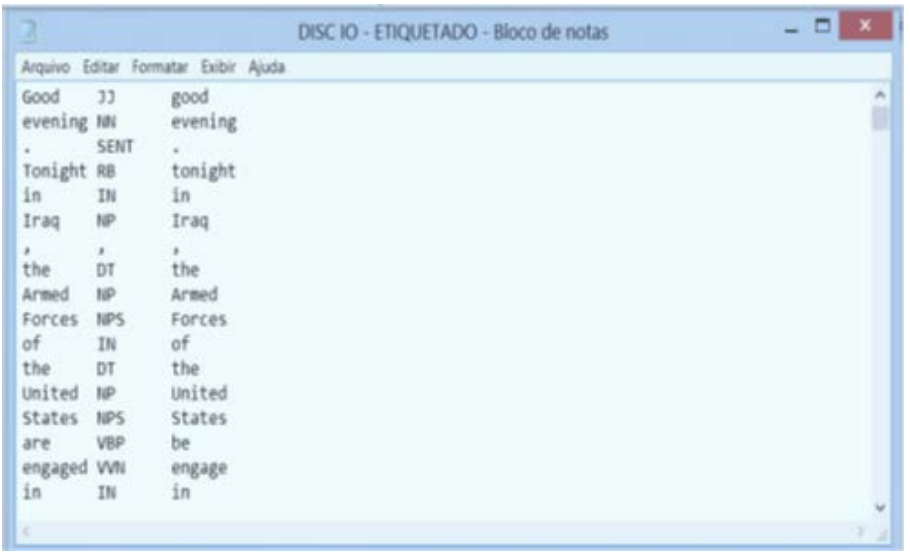
4. Procedimentos de anotação e análise automática

4.1. Etiquetamento morfossintático

O *software Treetagger*[5] foi desenvolvido por Helmud Schmid, linguista computacional da Universidade de Stuttgart, Alemanha. Trata-se de um anotador morfossintático capaz de etiquetar textos em formato eletrônico em diversos idiomas, dentre os quais incluem-se o inglês e o português. Como aponta Nunes (2014), para cada idioma, há um conjunto de documentos contendo parâmetros para a identificação automática de cada palavra (*type*) e/ou símbolo, a cada qual atribui-se uma classe gramatical e correspondente termo raiz (lema).

Cada texto do corpus (previamente salvo em formato *txt*) foi automaticamente etiquetado pelo programa. A Figura 2 apresenta uma amostra de um excerto de texto de discurso político do subcorpus IO, também gerado em arquivo de mesma extensão, após o processamento com a ferramenta.

FIGURA 2 – Excerto de texto anotado pela ferramenta *Treetagger*



Fonte: Nunes (2014, p. 84)

[5] Página com informações e instruções para *download* do programa: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/. Acesso em: 12 mar. 2019.

Como marca Nunes (2014), para cada língua, há um conjunto de etiquetas (*tagset*) processado pelo software para atribuição de categorias gramaticais para cada palavra em cada um dos textos dos quatro subcorpora do Klapt! O *tagset* para o inglês foi desenvolvido no projeto *Penn Treebank*,[6] idealizado e executado por pesquisadores dos departamentos de Linguística, Ciência da Computação e Ciência da Informação da Universidade da Pensilvânia, Estados Unidos. Já o conjunto de etiquetas das categorias morfossintáticas nos subcorpora de textos em português brasileiro foi desenvolvido pelo Grupo para o Processamento de Linguagem Natural (ProLNat@GE)[7] da Universidade de Santiago da Compostela, Espanha.

Nunes (2014) também menciona que a interface e os conjuntos de etiquetas têm como base as categorias da gramática tradicional em ambas as línguas, considerando-se a palavra como unidade de investigação.

Quanto ao grau de exatidão para ambos os *tagsets* reconhecidos pela interface, a probabilidade de correspondência palavra/símbolo-etiqueta varia entre 96 e 97% (SCHMID, 1994, p. 16).

Dadas as especificidades técnicas de cada *tagset*[8] em função das diferenças entre os sistemas linguísticos do inglês e do português, não há correspondência direta entre várias etiquetas nas duas línguas. Destarte, foi necessária elaboração manual de um parâmetro de equivalência entre elas de forma que apenas as categorias gramaticais comuns entre as referidas línguas pudessem ser contempladas para o processamento de dados. Pode-se assim somar dez classes de palavras partilhadas por ambas, quais sejam: adjetivo, advérbio, conjunção, determinante, interjeição, numeral, preposição, pronome, substantivo e verbo (NUNES, 2014).

O Quadro 1 apresenta o parâmetro criado para agrupar as etiquetas dos dois *tagsets*.

---

[6] Página eletrônica do projeto: https://catalog.ldc.upenn.edu/. Acesso em 12 mar. 2019.

[7] Endereço eletrônico do grupo: http://gramatica.usc.es/pln/index.html. Acesso em: 12 mar. 2019.

[8] As listas completas das etiquetas do inglês e do português podem ser visualizadas nos seguintes sítios eletrônicos: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html (inglês) / https://gramatica.usc.es/~gamallo/tagger.htm (português). Acesso em: 12 mar. 2019.

QUADRO 1 – Parâmetro de correspondência entre etiquetas dos *tagsets* do inglês e do português

Classe de palavra	Etiquetas do inglês	Etiquetas do português
Adjetivo	JJ, JJR, JJS	ADJ
Advérbio	RB, RBR, RBS, WRB	ADV
Conjunção	CC	CONJ
Determinante	DT	DET
Interjeição	UH	I
Numeral (cardinal e ordinal)	CD	CARD
Preposição	IN, IN/that, TO	PRP PRP+DET
Pronome	PP, PP$, WDT, WP, WP$	P PR
Substantivo	FW, NN, NNS, NP, NPS	NOM
Verbo	MD, VB, VBD, VBG, VBN, VBP, VBZ, VH, VHD, VHG, VHN, VHP, VHZ, VV, VVD, VVG, VVN, VVP, VVZ	V V+P

Fonte: Nunes (2014, p. 89)

Como pode-se visualizar no Quadro 1, os *tagsets* em inglês e português automaticamente processados pelo *TreeTagger* respectivamente atribuem às marcas conjuntivas as etiquetas CC e CONJ.

4.2 Processamento e extração automática de dados

Segundo Nunes (2014), uma vez identificadas as dez classes gramaticais comuns entre o inglês e o português e estabelecidas as correspondências entre as etiquetas dos respectivos *tagsets*, criou-se uma sequência de comandos para processamento e extração de dados quantitativos baseada nessas combinações.

Esse parâmetro compreende uma sequência de comandos (*script*)[9] e foi utilizado para: 1) extração da frequência de ocorrência de itens correspondentes às dez classes gramaticais, as quais incluem conjunções; 2) aplicação de testes de significância estatística para a frequência de ocorrência das conjunções. Este script foi esquematizado para ser processado pelo ambiente R, apresentado na próxima subseção.

4.2.1 O ambiente de programação R e o parâmetro para extração de dados

Entendido enquanto uma linguagem computacional livre,[10] o ambiente R pode, dentre inúmeros fins, ser utilizado para se extrair e manipular dados estatísticos, conforme assinala Nunes (2014). Esse ambiente foi desenvolvido por pesquisadores do departamento de Estatística da Universidade de Auckland, Nova Zelândia, e vem sendo paulatinamente aprimorado por contribuições provenientes de várias instituições de pesquisa. Trata-se de uma eficiente ferramenta de processamento e extração de dados numéricos e categóricos, sendo capaz de processá-los a partir de vários modelos estatísticos. Também permite aplicar testes de significância e executar funções dos mais variados tipos e graus de complexidade.

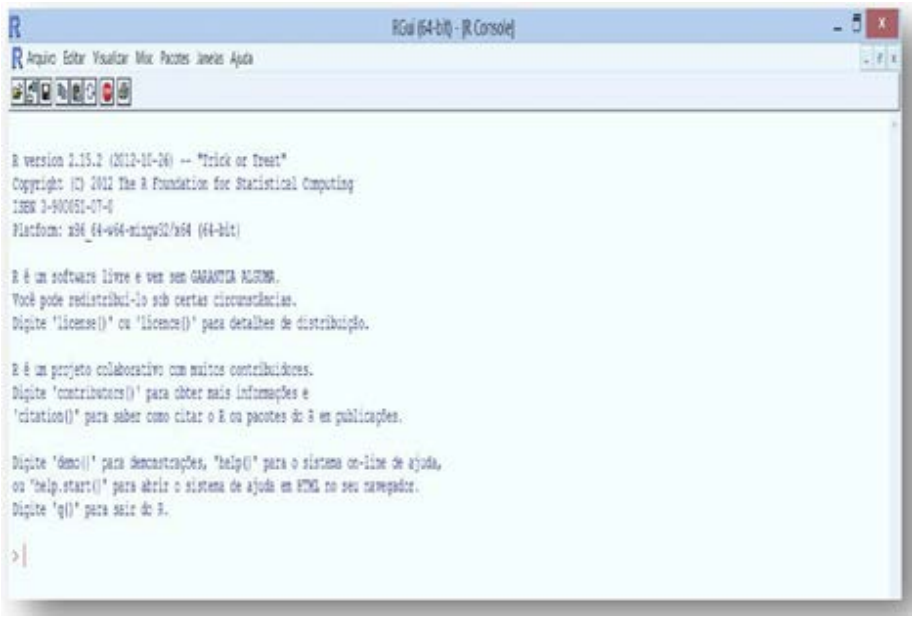
A Figura 3 apresenta a tela de exibição inicial do ambiente R.

---

[9] A sequência de comandos para a leitura automática dos arquivos, levantamento dos dados quantitativos e aplicação dos testes estatísticos está disponível na íntegra nos Anexos.

[10] Página do projeto R: http://www.r-project.org. Acesso em: 12 mar. 2018

FIGURA 3 – Tela inicial do ambiente R



Fonte: Nunes (2014, p. 92)

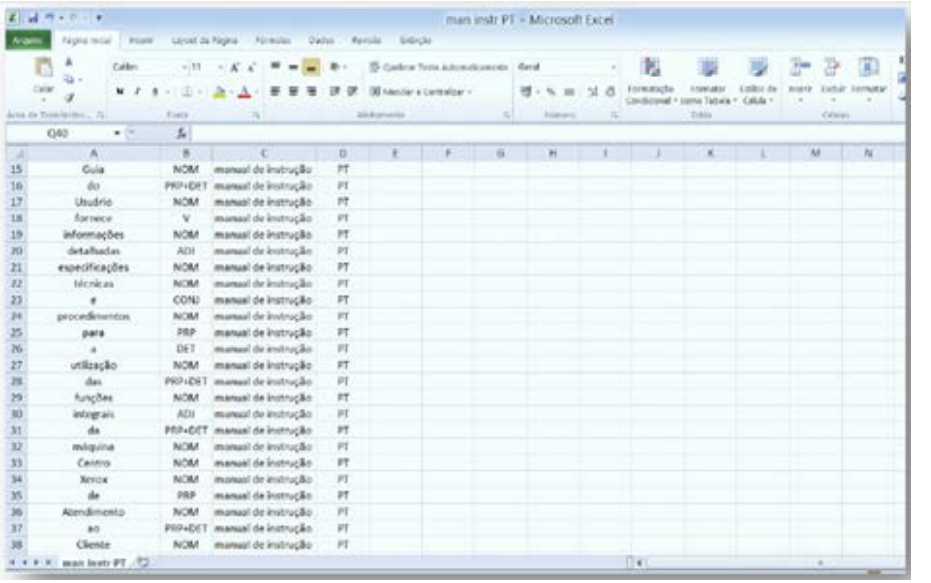
Para a extração dos dados, foi necessária manipulação prévia dos arquivos a serem processados pelo ambiente, bem como o desenvolvimento de um parâmetro com comandos para serem executados automaticamente.

Conforme descreve Nunes (2014), os arquivos contendo os textos de cada um dos quatro subcorpora foram agrupados por tipo, somando assim 32 (8 tipos textuais x 4 subcorpora). Em virtude de restrições técnicas de identificação de caracteres em textos com extensão *txt*, cada um destes arquivos foi convertido em planilhas eletrônicas do programa *Microsoft Excel©*.

Em cada planilha, foi suprimida a coluna contendo o lema de cada palavra etiquetada, já que esse elemento não figurou como objeto de análise. Em substituição a essa, duas outras colunas foram criadas: uma contendo o rótulo do respectivo tipo textual e outra explicitando o subcorpus (IO, PT, PO ou IT) correspondente.

A Figura 4 apresenta um exemplo de planilha eletrônica configurada no programa *Microsoft Excel©*.

FIGURA 4 – Configuração de planilha eletrônica para processamento no ambiente de programação R



Fonte: Nunes (2014, p. 93)

Formatados os 32 arquivos, criou-se manualmente um *script* para leitura e processamento desses pelo R. Esta sequência de comandos não apenas teve como função extrair, por subcorpus e tipo textual, a frequência absoluta total de itens de cada uma das dez classes gramaticais, mas também perfazer testes de significância estatística somente para a frequência de ocorrência das conjunções.

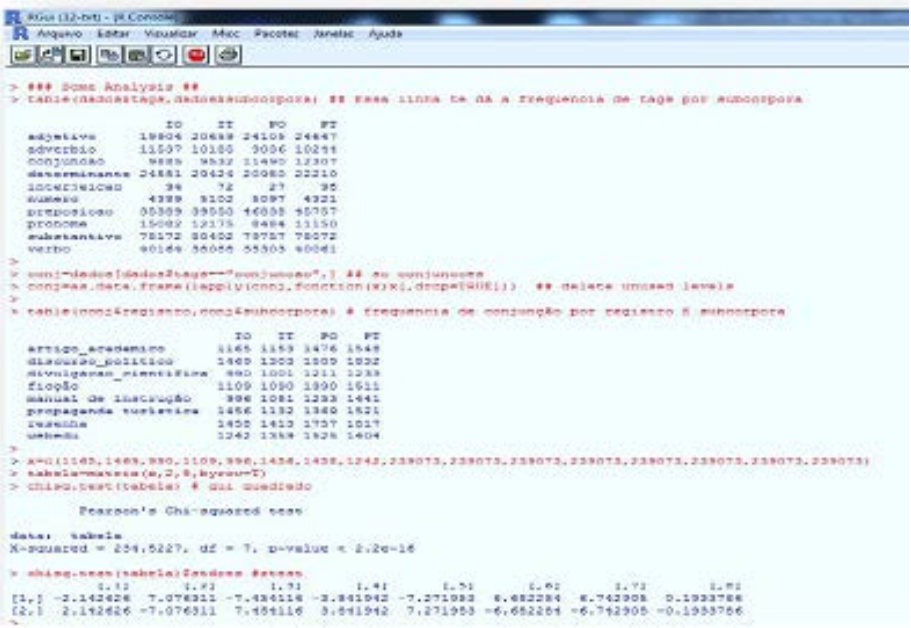
O parâmetro foi desenhado de forma que os dados pudessem ser processados conforme a seguinte ordem, conforme sequência descrita em Nunes (2014):

- Reconhecimento das 32 planilhas do *Microsoft Excel*©;
- Leitura e extração dos dados de cada planilha a partir do reconhecimento de cada palavra (*token*) e sua correspondente etiqueta morfossintática, bem como seu referido tipo textual e subcorpus;
- Correspondência das etiquetas morfossintáticas entre os *tagsets* do inglês e do português, e agrupamento em dez classes gramaticais, conforme parâmetro apresentado no Quadro 1;

– Extração da frequência absoluta de palavras (*tokens*) de cada classe gramatical por tipo textual e por subcorpus;
– Extração da frequência absoluta da classe gramatical *conjunção* por tipo textual e por subcorpus;[11]
– Aplicação do teste de significância estatística Qui-quadrado de aderência (*goodness-of-fit*) de Pearson a partir da frequência de ocorrência absoluta total da classe gramatical *conjunção* por subcorpus;
– Aplicação do teste de significância estatística *post hoc* Z a partir das frequências absolutas da classe de palavra *conjunção* por tipo textual (distribuídas nos 4 subcorpora).

A Figura 5 mostra uma representação do resultado (*output*) de parte dos dados gerados pelo *script* desenhado no estudo de Nunes (2014).

FIGURA 5 – Representação dos resultados gerados na interface do ambiente de programação R



Fonte: Nunes (2014, p. 95)

[11] Por questões de ordem pragmática, apenas as frequências de ocorrência desta classe de palavra estão apresentadas na seção de resultados e discussões.

Como destaca Nunes (2014), o teste Qui-quadrado de aderência (*goodness-of-fit*) e o teste *post hoc* Z verificaram se as distribuições das frequências absolutas das ocorrências da classe de palavra *conjunção* foram ou não estatisticamente significativas em cada tipo de texto e subcorpus do Klapt!. Foi possível averiguar, com o primeiro teste, se houve ou não desvios significativos da frequência de ocorrência geral esperada para as conjunções em cada subcorpus[12] ou se a frequência observada esteve dentro do previsto. Já o segundo teste revelou se as frequências de conjunções em cada tipo textual e subcorpus se mostraram significativamente acima ou abaixo das esperadas.

Ambos os testes figuram como ferramentas decisivas para verificação de hipóteses e/ou pressupostos por ventura aventados sobre a frequência de ocorrência de palavras e suas correspondentes classes gramaticais. Para o caso das conjunções, foi possível averiguar, numa perspectiva interlinguística, em que medida a hipótese da explicitação (cf. BLUM-KULKA, 1986) pode ser confirmada nos subcorpora de textos traduzidos. Já no prisma da variabilidade de registro entre os tipos textuais, pode-se confirmar ou não pressupostos sobre uma maior frequência de ocorrência de marcas de coesão textual em textos de caráter argumentativo (cf. NEUMANN, 2008) e nos textos originais e traduzidos em português brasileiro[13] (cf. VIEIRA, 1984).

Descritos os procedimentos para extração e processamento automático de dados, estabelece-se na próxima subseção uma relação entre a metodologia utilizada em Nunes (2014) para o escrutínio da frequência de ocorrência de conjunções e a validade dos resultados por esta obtidos.

4.3. Investigação automática vs. investigação manual de elementos linguísticos

Conforme exposto na seção anterior, os procedimentos de análise de frequência do corpus da investigação de Nunes (2014) envolveram a anotação morfossintática automática para posterior processamento de dados através de um ambiente de programação. O autor fundamenta-se no trabalho de Matthiessen (2009), que estabelece uma relação entre a abrangência de 1) resultados obtidos automaticamente e de 2) achados

[12] Cabe ressaltar que esta frequência de ocorrência se deu em relação ao número total de palavras (*tokens*) por subcorpus.

[13] Como aponta a autora, o português, em comparação ao inglês, apresenta maior grau de especificidade e clareza ao salientar recursos de coesão.

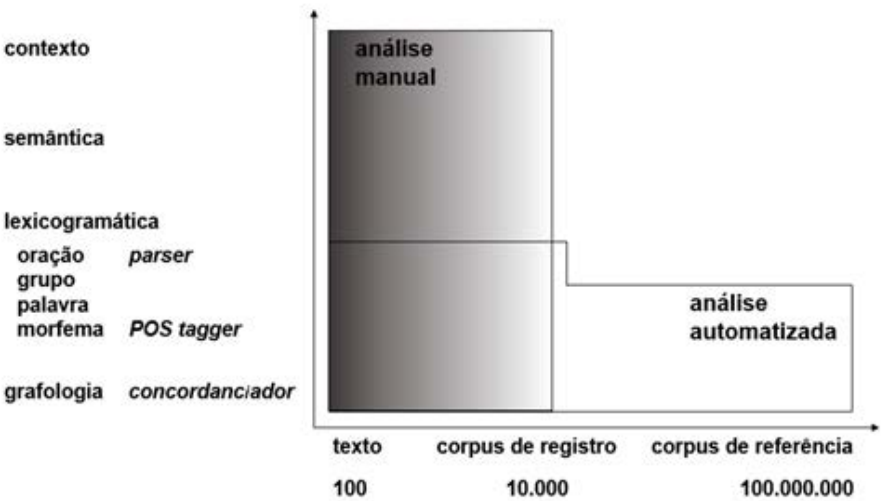
gerados manualmente em função da estratificação linguística e do tamanho do corpus.

De acordo com o Matthiessen, investigações em corpora de maiores extensões e que contemplam vários registros e tipos textuais podem ser facilmente realizadas automaticamente com o auxílio de ferramentas computacionais, o que soma para a descrição do potencial de construção de significados de determinada língua (MATTHIESSEN, 2009, p. 53 *apud* NUNES, 2014). Contudo, Matthiessen também reconhece a existência de restrições no escopo de análise semântica e contextual, já que as ferramentas de anotação, extração e processamento automático de um volume considerável de dados geralmente permitem somente o escrutínio lexicogramatical.

Já a investigação manual, geralmente realizada em corpora de menores extensões, permite maior aprofundamento de análise no que tange todos os níveis linguísticos. Em contrapartida, em virtude do tamanho reduzido das amostras e da variabilidade entre os registros e tipos textuais, há limitações em termos de descrição linguística.

Nunes (2014) ilustra essa relação marcada por Matthiessen (2009).

GRÁFICO 1 – Escopo da investigação manual e da investigação automática em relação aos níveis linguísticos e às dimensões do corpus.



Fonte: Matthiessen (2009, p. 53). Traduzido e adaptado por Nunes (2014)

Ao visualizar o Gráfico 1, pode-se afirmar que os procedimentos metodológicos de investigação de conjunções do corpus Klapt! descritos em Nunes (2014) produzem resultados que se situam na interseção entre a porção central do eixo horizontal (análise automatizada) e na porção inferior do eixo vertical (análise manual) da representação. Isso porque os insumos para a descrição linguística gerados por esses procedimentos são razoavelmente abrangentes em função do considerável tamanho e da variabilidade de registros observada no corpus como um todo. Todavia, os subsídios metodológicos resultam em achados que se restringem apenas às unidades lexicais dos textos, em razão de limitações impostas pelas ferramentas computacionais.

A próxima seção apresenta os principais resultados do estudo de Nunes (2014), bem como algumas discussões que emergiram a partir da apreciação geral desses achados.

5. Resultados e discussões

Conforme exposto anteriormente, os procedimentos de anotação e extração automática das frequências de ocorrência da classe gramatical *conjunção* na investigação de Nunes (2014) produziram resultados que apontaram em qual (ou quais) dos oito tipos de textos do corpus paralelo bidirecional bilíngue as frequências observadas das marcas conjuntivas foram estatisticamente significativas. Ainda, esses resultados permitiram testar hipóteses e pressupostos sobre o impacto da variabilidade de registro (entre os tipos de textos) e da tipologia linguística na frequência de ocorrência desses elementos linguísticos no corpus em questão.

As frequências absolutas de conjunções distribuídas por tipo textual e por subcorpus do Klapt! se encontram apresentadas na Tabela 2.

TABELA 2 – Frequência absoluta de conjunções no Klapt! por subcorpus e tipo textual

Subcorpus / Tipo textual	Inglês original (IO)	Inglês traduzido (IT)	Português original (PO)	Português traduzido (PT)	Total por tipo textual
Artigo acadêmico	1.165	1.153	1.476	1.548	**5.342**
Discurso político	1.469	1.303	1.509	1.832	**6.113**
Divulgação científica	990	1.001	1.211	1.233	**4.435**
Ficção	1.109	1.090	1.390	1.511	**5.100**
Manual de instrução	996	1.081	1.253	1.441	**4.771**
Propaganda Turística	1.456	1.132	1.369	1.521	**5.478**
Resenha	1.458	1.413	1.757	1.817	**6.445**
Website Educacional	1.242	1.359	1.525	1.404	**5.530**
Total por subcorpus	**9.885**	**9.532**	**11.490**	**12.307**	

Fonte: Nunes (2014, p. 110)

Os resultados de Nunes (2014) explicitam que, numa perspectiva interlinguística, houve uma inversão proporcional nas frequências absolutas de conjunções nos dois subcorpora paralelos. Enquanto existe um número superior dessas marcas em todos os tipos textuais do subcorpus PT se comparado a todos os mesmos tipos textuais no subcorpus IO, há um número inferior de conjunções no subcorpus IT se comparado ao subcorpus PO. Nunes também aponta que tal inversão a princípio poderia sugerir não ser integralmente corroborada a hipótese de que a frequência de conjunções se mostraria acima da esperada nos subcorpora de textos traduzidos (partindo-se da hipótese de explicitação de elementos coesivos (cf. BLUM-KULKA, 1986). Em contrapartida, Nunes também assinala que a hipótese de que a frequência de conjunções nos textos em português brasileiro estaria dentro ou acima da esperada poderia ser confirmada com base no pressuposto que a língua portuguesa (comparando-se ao inglês) evidencia "maior grau de especificidade e clareza ao evidenciar recursos coesivos" (cf. VIEIRA, 1984 *apud* NUNES, 2014).

Observando-se pelo prisma da variabilidade de registro entre os tipos textuais, as distribuições revelam maior frequência de ocorrência de conjunções nos textos de resenhas e de discursos políticos. Em uma primeira instância, ambos esses números corroboram o pressuposto de Neumann (2008), a qual advoga que "textos argumentativos tendem a tornar expressas conexões entre as porções textuais para que a coerência textual seja estabelecida" (NEUMANN, 2008, p. 109 *apud* NUNES, 2014).

Dado que as frequências apresentadas na Tabela 2 são absolutas, fez-se necessário testes de significância para se confirmar ou refutar as hipóteses e os pressupostos fundamentados nas proposições das referidas autoras. Para tal, verificou-se se as distribuições das frequências apresentadas para cada subcorpus e tipo textual foram ou não significativas via aplicação de testes estatísticos (NUNES, 2014).

Conforme explicado na subseção 3.2.1, o estudo de Nunes (2014) se valeu do teste do Qui-quadrado de Pearson, aplicado para se averiguar se a frequência total de conjunções observada em cada subcorpus se desviou ou não da frequência esperada em relação ao número total de palavras de cada subcorpus. A tabela a seguir mostra os resultados desse teste.

TABELA 3 – Resultados do teste do Qui-quadrado para a frequência de ocorrência de conjunções por subcorpus

Subcorpus	**Resultado**
Inglês original (IO)	$X^2 = 234,5227$, df = 7, p-value < 0,01
Inglês traduzido (IT)	$X^2 = 128,0775$, df = 7, p-value < 0,01
Português original (PO)	$X^2 = 144,3397$, df = 7, p-value < 0,01
Português traduzido (PT)	$X^2 = 184,5895$, df = 7, p-value < 0,01

Fonte: Nunes (2014, p. 111)

Os resultados na tabela revelam um desvio significativo entre a frequência observada e a frequência esperada em cada um dos quatro subcorpora, uma vez que o valor de p (*p-value*) foi consideravelmente inferior a 0,05[14] em todos eles. Como afirma Nunes (2014, p. 112), foi

[14] Na esfera dos estudos linguísticos e das ciências humanas, este valor é parâmetro para significância estatística. (GRIES, 2012j *apud* NUNES, 2014)

"extremamente baixa a probabilidade de essa frequência ter acontecido ao acaso considerando-se uma distribuição uniforme para as conjunções em cada subcorpus".

Para se localizar os desvios significativos em cada tipo de texto de cada subcorpus, aplicou-se o teste *post hoc* Z[15] e observou-se os resultados em seis perspectivas, respectivamente: duas paralelas (IO-PT e PO-IT), duas comparáveis monolíngues (IO-IT e PO-PT) e duas comparáveis bilíngues (IO-PO e IT-PT).

Os pares de escores Z,[16] distribuídos nas duas perspectivas paralelas e apresentados na Tabela 4, revelam a medida em que as duas línguas (inglês e português brasileiro) originais e traduzidas impactaram na frequência de ocorrência de conjunções no corpus como um todo. Também mostram o impacto da variabilidade de registro entre os tipos textuais na frequência desses elementos linguísticos.

---

[15] Cada número nas tabelas corresponde a um escore Z, ou seja, o resíduo (variação) acima ou abaixo de uma frequência considerada esperada (equivalente ao número 0). Os números positivos distintos de 0 estão acima de uma distribuição esperada e os negativos estão abaixo. Para existir significância estatística, os escores positivos devem ser iguais ou superiores a 1,96 e os negativos iguais ou inferiores a -1,96 (BARONI; EVERT, 2008, p. 13 *apud* NUNES, 2014). Nunes (2014) ainda afirma que se um escore atingir qualquer um desses dois parâmetros mínimos de significância (um positivo e outro negativo), a probabilidade de determinada frequência ter se desviado da frequência esperada de forma fortuita é muito baixa. Por outro lado, se os valores estiverem entre esses dois polos (-1,96 e 1,96), há probabilidade de que a frequência de ocorrência de conjunções tenha acontecido por acaso.

[16] No estudo de Nunes (2014), foram destacados tanto os pares de distribuições mais estatisticamente significativos nas duas perspectivas paralelas e nas quatro perspectivas comparáveis, cujas magnitudes dos escores Z apontaram maiores diferenças entre si. O autor também explicita que as diferenças foram evidenciadas pelos valores que marcam uma oposição entre 1) duas distribuições (uma estatisticamente significativa acima da esperada e outra significativamente abaixo da esperada, ou vice-versa) ou 2) entre uma distribuição não estatisticamente significativa e outra estatisticamente significativa. Os pares de escores mais relevantes estão em negrito em todas as seis perspectivas de análise.

TABELA 4 – Distribuições do teste *post hoc* Z para a frequência de conjunções nos subcorpora paralelos

Subcorpus / **Tipo textual**	**Inglês original (IO) – Português brasileiro traduzido (PT)**	**Português original (PO) – Inglês traduzido (IT)**
Artigo acadêmico	**-2,142626 / 0,261529**	1,117886 / -1,189558
Discurso político	7,076311 / 7,974455	2,045824 / 3,444186
Divulgação científica	-7,454116 / -8,302080	-6,337658 / -5,887555
Ficção	**-3,841942 / -0,74388**	**-1,300883 / -3,136453**
Manual de instrução	-7,271953 / -2,646343	-5,155558 / -3,414615
Propaganda Turística	**6,682284 / -0,47213**	-1,891624 / -1,838476
Resenha	6,742905 / 7,567270	9,015954 / 6,840729
Website Educacional	**0,193379 / -3,652116**	2,495694 / 5,173498

Fonte: Nunes (2014, p. 114)

Levando-se em consideração a variabilidade de registro entre as distribuições das frequências nos subcorpora paralelos, os escores contidos na Tabela 4 corroboram a hipótese de que tipos textuais de caráter argumentativo (como textos dos tipos *discurso político* e *resenha*) tendem a apresentar frequências de conjunções significativamente acima das esperadas,[17] já que esses elementos linguísticos compreendem "um recurso coesivo retórico que comumente ocorre em textos dessa natureza quando comparados a textos de outros tipos" (cf. NEUMANN, 2008 *apud* NUNES, 2014).

Pelo prisma da tipologia linguística, o autor advoga que as distribuições das perspectivas paralelas apresentadas na Tabela 4 vão de encontro à hipótese de que a frequência de ocorrência de conjunções nos textos traduzidos seria significativamente acima da esperada se comparada à frequência aos seus respectivos textos originais. Nunes argumenta, assim, que não se pode generalizar a hipótese fundamentada no fenômeno da explicitação (cf. BLUM-KULKA, 1986) em textos

[17] Vale ressaltar que esses resultados se repetiram nas perspectivas comparáveis mono e bilíngues.

traduzidos. Isso porque, segundo ele, nem todos os escores Z das frequências nos textos traduzidos foram significativamente acima dos esperados comparando-se às distribuições dos seus respectivos textos originais. Ele ainda destaca que, ao contrário do que se poderia esperar, alguns tipos de texto (como *ficção* e *website educacional*, por exemplo) apresentaram distribuições significativamente abaixo das esperadas nos textos traduzidos em inglês e em português brasileiro quando cotejados com seus respectivos textos originais.

Nunes (2014) ainda mostra que, ao se comparar os valores de cada par de escores na direção inglês-português brasileiro, puderam ser notadas discrepâncias salientes nas distribuições das frequências nos tipos *artigo acadêmico*, *ficção*, *propaganda turística* e *website educacional,* com destaque para os dois últimos. O autor demonstra que nos primeiros dois tipos de texto, enquanto ambas as frequências se mostraram significativamente abaixo das esperadas nos textos originais, suas traduções apresentaram, respectivamente, frequência abaixo e acima da esperada, porém sem significância estatística. No tipo textual *propaganda turística*, a frequência de conjunções nos originais foi expressivamente acima da esperada, ao passo que nas traduções ela se mostrou abaixo da esperada (contudo sem significância estatística nesta última). Já no tipo textual *website educacional,* enquanto a distribuição obtida para os textos originais foi pouco acima da esperada, nos textos traduzidos esta frequência se mostrou consideravelmente abaixo da esperada.

Na direção português brasileiro-inglês, a comparação entre os valores se fez relevante apenas em textos do tipo *ficção*, cuja frequência nos originais esteve abaixo da esperada e, nas traduções, significativamente abaixo da esperada.

Os resultados de Nunes (2014) para os subcorpora paralelos foram produtivos sobretudo no que toca os textos do tipo *ficção* ao mostrarem nestes 1) uma tendência de se explicitar conjunções em português brasileiro traduzidos do inglês e 2) uma tendência de não se explicitar esses recursos coesivos em textos em inglês traduzidos do português brasileiro.

Conforme sinalizado anteriormente, os pares de escores Z também foram observados pelas duas perspectivas comparáveis no referido estudo. A Tabela 5 apresenta as distribuições nos dois subcorpora comparáveis monolíngues.

TABELA 5 – Distribuições do teste Z para frequência de conjunções nos subcorpora comparáveis monolíngues

Subcorpus / **Tipo textual**	**Inglês original (IO) – Inglês traduzido (IT)**	**Português original (PO) – Português traduzido (PT)**
Artigo acadêmico	**-2,142626 / -1,189558**	1,117886 / 0,2615289
Discurso político	7,076311 / 3,444186	2,045824 / 7,974455
Divulgação científica	-7,454116 / -5,887555	-6,337658 / -8,302080
Ficção	-3,841942 / -3,136453	-1,300883 / -0,7438762
Manual de instrução	-7,271953 / -3,414615	-5,155558 / -2,646343
Propaganda Turística	**6,682284 / -1,838476**	-1,891624 / -0,4721325
Resenha	6,742905 / 6,840729	9,015954 / 7,567270
Website Educacional	**0,193379 / 5,173498**	**2,495694 / -3,652116**

Fonte: Nunes (2014, p. 115)

Os escores dos subcorpora de textos em inglês originais e traduzidos em Nunes (2014) revelaram maior produtividade para os tipos textuais *artigo acadêmico*, *propaganda turística* e *website educacional,* sendo que estes últimos dois tipos apresentam as comparações mais significativas. Ambas as distribuições, para o primeiro tipo textual, apresentam valores de frequência abaixo da esperada. Porém, estas se mostraram estatisticamente significativas apenas nos textos originais. Já os textos de propaganda turística apresentaram escores consideravelmente acima dos esperados nos textos originais e abaixo dos esperados (sem significância estatística) nos textos traduzidos. O tipo textual *website educacional*, por sua vez, apresentou nos textos originais escore acima dos esperados (porém sem significância estatística) e escore significativamente acima dos esperados nos textos traduzidos

As distribuições nos textos originais e traduzidos em português brasileiro, por sua vez, ratificaram parcialmente a hipótese de que a as frequências das conjunções nestes, em comparação aos textos originais e traduzidos para o inglês, estariam dentro ou acima da frequência esperada, considerando-se a sugestão de que "os recursos coesivos naquela língua refletem maior grau de clareza e especificidade do que nesta" (cf.

VIEIRA, 1984 *apud* NUNES, 2014). Ainda, notou-se que, assim como nos textos originais e traduzidos para o inglês, as distribuições dos escores Z para os textos de *divulgação científica* e *manuais de instrução* no subcorpus comparável monolíngue em português brasileiro se encontram significativamente abaixo da frequência esperada para estes tipos textuais.

Os resultados do subcorpus monolíngue em português brasileiro ainda indicaram que o tipo *website educacional* apresentou maior relevo, já que a distribuição da frequência de conjunções foi significativamente acima da esperada nos textos originais e significativamente abaixo nos textos traduzidos, indicando assim uma oposição entre as magnitudes.

Se na perspectiva paralela a tipologia linguística impactou na frequência de conjunções, sobremaneira em textos de ficção, na perspectiva comparável monolíngue tal tipologia (língua original *versus* língua traduzida neste par linguístico) também se mostrou determinante na frequência destes recursos coesivos no tipo *website educacional* em ambos os subcorpora.

As distribuições das frequências nos subcorpora comparáveis bilíngues estão apresentadas na Tabela 6.

TABELA 6 - Distribuições do teste Z para frequência de conjunções nos subcorpora comparáveis bilíngues

Subcorpus / **Tipo textual**	**Inglês original (IO) – Português original (PO)**	**Inglês traduzido (IT) – Português traduzido (PT)**
Artigo acadêmico	**-2,142626 / 1,117886**	-1,189558 / 0,2615289
Discurso político	7,076311 / 2,045824	3,444186 / 7,974455
Divulgação científica	-7,454116 / -6,337658	-5,887555 / -8,302080
Ficção	**-3,841942 / -1,300883**	**-3,136453 / -0,7438762**
Manual de instrução	-7,271953 / -5,155558	-3,414615 / -2,646343
Propaganda Turística	**6,682284 / -1,891624**	-1,838476 / -0,4721325
Resenha	6,742905 / 9,015954	6,840729 / 7,567270
Website Educacional	**0,193379 / 2,495694**	**5,173498 / -3,652116**

Fonte: Nunes (2014, p. 117)

As distribuições da tabela não ratificam totalmente a hipótese baseada no pressuposto de que, em virtude de explicitação de recursos coesivos em línguas alvo (cf. BLUM-KULKA, 1986), os textos traduzidos em inglês e em português brasileiro apresentariam frequência significativamente acima da esperada se comparada às frequências dos textos originais nessas línguas. Nesta perspectiva de análise, Nunes assinala que, em algumas distribuições dos subcorpora de textos traduzidos, as frequências observadas se mostraram significativamente a) abaixo das frequências esperadas (como em textos do tipo *website educacional* e *divulgação científica* em português brasileiro) e b) significativamente acima das frequências esperadas, porém com valor inferior às frequências observadas nos textos originais (como se observa nas resenhas traduzidas em português brasileiro).

Quanto à comparação entre as distribuições nos subcorpora dos textos originais, Nunes também aponta que houve significância para os seguintes tipos textuais: *artigo acadêmico*, *ficção, propaganda turística* e *website educacional*. Para o primeiro tipo, os números revelaram que as frequências nos textos em inglês se mostraram significativamente abaixo das esperadas, ao passo que estiveram acima das esperadas as frequências nos textos em português brasileiro, porém sem significância estatística. Nos textos de ficção, ambas as frequências se mostraram abaixo das esperadas, mas houve relevância estatística apenas nos textos originais em português brasileiro. A frequência nos textos de propagandas turísticas originais em inglês, por sua vez, se mostrou significativamente acima da esperada, ao passo que, nos textos originais em português do mesmo tipo, a frequência esteve abaixo da esperada (porém sem significância estatística para esta última). Por fim, a distribuição da frequência em textos de websites educacionais originais em inglês se mostrou acima da esperada (sem relevância estatística) e significativamente acima da esperada nos textos originais em português brasileiro.

Já no subcorpus de textos traduzidos, os tipos *ficção* e *website educacional* apresentaram maior relevo na comparação de ambas as distribuições. Para o primeiro tipo textual, ambas as frequências se mostraram abaixo das esperadas, sendo que nos textos traduzidos para o inglês a frequência se mostrou significativamente abaixo da esperada e nos textos traduzidos para o português a frequência se mostrou abaixo da esperada (todavia sem significância estatística). Já no tipo *website educacional,* a distribuição da frequência nos textos traduzidos para o

inglês se mostrou consideravelmente acima da esperada, estabelecendo assim uma oposição com a frequência dos textos traduzidos para o português brasileiro, que se mostrou significativamente abaixo da esperada.

Em termos gerais, os resultados das análises de Nunes (2014) nas duas perspectivas paralelas e das quatro comparáveis revelaram que: 1) a variabilidade funcional de registro se mostrou como a variável com maior impacto na frequência de ocorrência de conjunções no corpus, em detrimento da tipologia linguística. Isso pode ser comprovado pela hipótese baseada em Neumann (2008) ratificada na frequência dessas marcas em textos com caráter argumentativo (*resenha e discurso político*); 2) os pressupostos com base em Vieira (1984) não foram integralmente corroborados, tendo em vista que nem todas as frequências nos textos em português (originais e traduzidos) se mostraram dentro ou acima das esperadas; 3) a hipótese baseada em Blum-Kulka (1986) foi parcialmente corroborada, uma vez que nem todas as frequências nos textos traduzidos nas duas línguas se mostraram acima das frequências esperadas em comparação aos seus respectivos textos originais e/ou aos seus respectivos textos comparáveis na mesma língua.

Os resultados da referida investigação ainda apontaram relevância na frequência de ocorrência em dois tipos de texto: *ficção* e *website educacional*. O primeiro apresentou significância das frequências nas perspectivas paralelas e comparáveis bilíngues, revelando, em suma, uma menor tendência de explicitação de conjunções neste tipo de texto em língua inglesa original e traduzida (do português brasileiro) em comparação com textos do mesmo tipo em português brasileiro. O segundo tipo, por sua vez, revelou frequências significativas nas perspectivas paralelas e comparáveis mono e bilíngues, sugerindo uma tendência de menor explicitação de conjunções em língua portuguesa brasileira traduzida do inglês, e de maior explicitação dessas marcas em língua inglesa traduzida do português brasileiro.

A próxima seção sintetiza os resultados apresentados por Nunes (2014), além de tecer suas implicações sobretudo para o campo disciplinar dos estudos da tradução.

6. Conclusões

O estudo de Nunes (2014) apresentou achados de análise automática de elementos de coesão obtidos a partir de etiquetamento

morfossintático e processamento automático de dados. Foram extraídas frequências de ocorrência de conjunções em corpus paralelo bidirecional bilíngue no par linguístico inglês-português brasileiro composto por textos de oito tipos distintos. Os resultados destas frequências foram analisados tanto pela perspectiva da tipologia linguística quanto pelo viés da variabilidade de registro entre os tipos textuais.

Na esfera dos estudos descritivos da tradução orientados ao produto, pode-se concluir que os achados da análise automática da frequência de conjunções podem auxiliar na descrição linguística do inglês e da variante brasileira do português em seus modos escritos, já que destacaram os diferentes potenciais de explicitação de conjunções de acordo com o tipo textual e com a tipologia texto original *versus* texto traduzido. Como apontaram os resultados, textos ficcionais e de websites educacionais foram os tipos mais produtivos dessa investigação: nestes, evidenciou-se tendência à explicitação de conjunções nos textos traduzidos em inglês, e, naqueles, observou-se esta mesma tendência nos textos originais e traduzidos em português brasileiro.

No âmbito metodológico, o estudo se mostrou relevante no sentido de apresentar ferramentas produtivas de atribuição de categorias gramaticais (tendo a palavra como unidade de análise) para a investigação de textos originais e traduzidos. Ainda, descreveu, de forma sequenciada, procedimentos de extração automática e tratamento estatístico de dados para a geração de resultados com maior confiabilidade.

Os achados da pesquisa ainda ensejam potencial de aplicação em sistemas de tradução automática e como subsídio pedagógico na formação de tradutores em tarefas que envolvam o par linguístico inglês-português brasileiro e em ambas as direções.

Como principal apontamento para pesquisas futuras está a replicação da metodologia aqui descrita para a investigação, com aporte estatístico, da frequência de outras classes gramaticais em corpora bidirecionais paralelos bilíngues.

Agradecimentos

O autor agradece os esforços de toda a equipe de docentes e discentes que compõe o Laboratório Experimental de Tradução (LETRA), ao André Souza, pelo auxílio no tratamento estatístico dos dados, e também à CAPES e ao CNPq pelo fomento financeiro.

Referências

BARONI, Marco; EVERT, Stefan. Statistical methods for corpus exploitation. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 2008. Chapter 38.

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

BLUM-KULKA, Shoshana. Shifts of cohesion and coherence in translation. *In*: HOUSE, Juliane; BLUM-KULKA, Shoshana (ed.). *Interlingual and Intercultural Communication*. Tübingen: Narr, 1986. p. 17-35.

GRANGER, Sylviane. The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies? *In*: GRANGER, Sylviane; LEROT, Jacques; PETCH-TYSON, Stephanie (ed.). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam; New York: Rodopi, 2003. p. 17-29.

GRIES, Stefan. Testing independent relationships. *In*: CHAPELLE, Carol A. (ed.). *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell, 2012. Doi: https://doi.org/10.1002/9781405198431.wbeal1202

HALLIDAY, M.A.K.; HASAN, R. *Cohesion in English*. Essex: Longman Group UK Limited, 1976.

HALLIDAY, M. A. K.; MATTHIESSEN, Christian M. I. M. *Halliday's Introduction to Functional Grammar*. 4th ed. Oxon; New York: Routledge, 2014.

HOLMES, James S. The Name and Nature of Translation Studies. *Third International Congress of Applied Linguistics*. Copenhagen: [*S.n.*], 1972.

JESUS, Silvana M.; NUNES, Leonardo P. Klapt! Corpus Design and Compilation. *In*: KUNZ, Kerstin; TEICH, Elke; HANSEN-SCHIRRA, Silvia; NEUMANN, Stella; DAUT, Peggy (org.). *Caught in the Middle:* Language Use and Translation. A Festschrift for Erich Steiner on the Occasion of his 60th Birthday. Saarbrücken: Universitätsverlag des Saarlandes, 2014. p. 177-193.

MATTHIESSEN, Christian M. I. M; Ideas and new directions. In: HALLIDAY, Michael A. K; WEBSTER, Jonathan J. (ed.). *Continuum Companion to Systemic Functional Linguistics*. London; New York: Continuum International Publishing Group, 2009.

MATTHIESSEN, Christian M. I. M.; TERUYA, Kazuhiro; LAM, Marvin. *Key Terms in Systemic Functional Linguistics*. London; New York: Continuum International Publishing Group, 2010.

NEUMANN, Stella. Corpus Design. In: *Linguistic properties of translations: a corpus based investigation for the language pair English-German.* Deliverable no.1 DFG project STE 840/5-1, 2005. Disponível em: http://fedora.clarin-d.uni-saarland.de/croco-gecco/croco/corpus_design.pdf. Acesso em: 12 mar. 2019

NEUMANN, Stella. *Contrastive Register Variation.* A Quantitative Approach to the Comparison of English and German. Habilitationsschrift. Saarbrücken: Philosophische Fakultät II, Universität des Saarlandes, 2008.

NEVES, M. H. M. *Gramática de usos do português*. 2. ed. São Paulo: Unesp, 2011.

NUNES, Leonardo P. *As conjunções* but *e* mas *em textos ficcionais originais e traduzidos*: uma abordagem tridimensional com base na linguística sistêmico-funcional. 2010. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

NUNES, Leonardo P. *Relações coesivas e estruturais*: um estudo de conjunções em corpus paralelo e comparável no par linguístico inglês – português brasileiro. 2014. Tese (Doutorado em Linguística Aplicada) – Programa de Pós-Graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2014.

SCHMID, Helmud. Probabilistic Part-of-Speech Tagging Using Decision Trees. *In*: INTERNATIONAL CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING, 1994, Manchester, UK. *Proceedings [...]*. Manchester, UK: [*S. n.*], 1994.

VIEIRA, Else P. R. Comparative Stylistics Applied to Translation from English into Portuguese. *In*: WORLD CONGRESS OF APPLIED LINGUISTICS, 7., 1984, Brussels, Belgium. *Proceedings [...]*. Brussels: AILA Brussels, 1984. v. 3, n.26-36, p. 1275-1276.