



UNIVERSIDADE DE BRASÍLIA
FACULDADE DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

THIAGO BLANCH PIRES

AMPLIANDO OLHARES SOBRE A TRADUÇÃO AUTOMÁTICA ONLINE: UM ESTUDO
EXPLORATÓRIO DE CATEGORIAS DE ERROS DE MÁQUINA DE TRADUÇÃO GERADOS
EM DOCUMENTOS MULTIMODAIS

Orientador: Prof. Dr. Cláudio Gottschalg-Duque

BRASÍLIA/DF

MARÇO/2017

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

AMPLIANDO OLHARES SOBRE A TRADUÇÃO AUTOMÁTICA ONLINE: UM ESTUDO
EXPLORATÓRIO DE CATEGORIAS DE ERROS DE MÁQUINA DE TRADUÇÃO GERADOS
EM DOCUMENTOS MULTIMODAIS

THIAGO BLANCH PIRES

Orientador: Prof. Dr. Cláudio Gottschalg-Duque

TESE DE DOUTORADO

BRASÍLIA, DF

MARÇO/2017

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

THIAGO BLANCH PIRES

AMPLIANDO OLHARES SOBRE A TRADUÇÃO AUTOMÁTICA ONLINE: UM ESTUDO
EXPLORATÓRIO DE CATEGORIAS DE ERROS DE MÁQUINA DE TRADUÇÃO GERADOS
EM DOCUMENTOS MULTIMODAIS

Tese de doutorado apresentada ao
Programa de Pós-Graduação em Ciência da
Informação, da Faculdade de Ciência da
Informação da Universidade de Brasília,
como requisito parcial à obtenção do título
de doutor em Ciência da Informação.

Orientador: Prof. Dr. Cláudio Gottschalg-
Duque

BRASÍLIA, DF

MARÇO/2017

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Pires, Thiago Blanch
PP667a Ampliando olhares sobre a tradução automática
online: um estudo exploratório de categorias de erros
de máquina de tradução gerados em documentos
multimodais. / Thiago Blanch Pires; orientador
Cláudio Gottschalg Duque. -- Brasília, 2017.
167 p.

Tese (Doutorado - Doutorado em Ciência da
Informação) -- Universidade de Brasília, 2017.

1. Tradução automática. 2. Multimodalidade. 3.
Avaliação de tradução automática. 4. Textura
intersemiótica. 5. Incompatibilidade intersemiótica.
I. Duque, Cláudio Gottschalg, orient. II. Título.




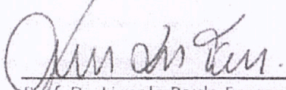
FOLHA DE APROVAÇÃO

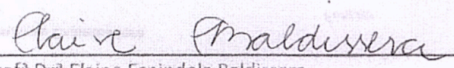
Título: "AMPLIANDO OLHARES SOBRE A TRADUÇÃO AUTOMÁTICA ONLINE: UM ESTUDO EXPLORATÓRIO DE CATEGORIAS DE ERROS DE MÁQUINA DE TRADUÇÃO GERADOS EM DOCUMENTOS MULTIMODAIS".

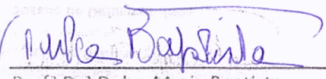
Autor (a): Thiago Blanch Pires
Área de concentração: Gestão da Informação
Linha de pesquisa: Organização da Informação

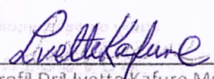
Tese submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Doutor** em Ciência da Informação.
Tese aprovada em: 02 março de 2017.


Prof. Dr. Cláudio Gottschalg-Duque
Presidente (UnB/PPGCINF)


Prof. Dr. Lincoln Paulo Fernandes
Membro Externo (UFSC)


Prof. Dr. Elaine Espindola Baldissera
Membro Externo (PUCV-Chile)


Prof. Dr. Dulce Maria Baptista
Membro Interno (UnB/PPGCINF)


Prof. Dr. Ivette Kafure Munoz
Membro Interno (UnB/FCI)

Prof. Dr. André Porto Ancona Lopez
Coordenador (UnB/PPGCINF)

Dedicado a Sandra, Donizete e Rafaela.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais Sandra e Donizete pelo apoio e acolhimento incondicional durante os estágios mais críticos do doutorado. À minha irmã Rafaela, que para além de todo apoio emocional também disponibilizou-se em dialogar, e organizar ideias sobre diversos pontos dessa pesquisa.

Ao professor Cláudio, agradeço imensamente pela abertura, incentivo, e acompanhamento para buscar caminhos inovadores na pesquisa acadêmica. Aos colegas do grupo de pesquisa R.E.G.I.I.M.E.N.T.O. agradeço pelas discussões instigantes e cooperações na organização dos eventos do grupo.

Não tenho palavras para agradecer a participação dos professores membros da minha banca de defesa: profa. dra. Elaine Espíndola e prof. dr. Lincoln Fernandes, que acompanham minha trajetória acadêmica desde o início, e aos professores membros internos Dulce Baptista, em especial, pelos ensinamentos em sua disciplina de Organização da Informação, profa. dra. Ivette Kafure e prof. dr. André Porto Lopez, pela contribuição acadêmica e por aceitarem prontamente o convite para participar da banca.

I'm grateful to have been supervised by professor John Bateman during my Phd visit at Universität Bremen. I thank all his support for developing my research and putting me in contact with everyone who contributed to my research visit. I thank all members of John's working group I had the chance to work with: Janina, Tuomo, Ognyan and Daniela for all the academic work, daily support and friendship. To all of them my most sincere gratitude for presenting me to the "Multimodality world".

To my friends during the 9-month stay in Bremen, Volkhard, and the "Bremer Stadtmusikanten", Lukrecia, Patricks and Helia, my most sincere gratitude.

Agradeço à Universidade de Brasília por ter concedido três anos de licença para a realização do doutorado. Agradeço também aos professores e colegas do departamento LET, mais especificamente do curso de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI), que aprovaram e apoiaram meu afastamento para realização do doutorado.

O aprofundamento de estudos no campo da multimodalidade, a experiência acadêmica e cultural compreendida entre abril de 2015 e dezembro do mesmo ano só foram possíveis por meio da concessão de bolsa do Programa de Bolsa de Doutorado Sanduíche (PDSE) da CAPES.

Por fim, mas não menos importante, agradeço aos amigos Rodrigo de Sales e Fernando Silva, pelas diversas conversas sobre multimodalidade, tradução automática, e ciência da informação; à amiga Alessandra, pela grande presença de espírito durante

as aulas, discussões e trabalhos acadêmicos; aos amigos Fernanda e Vinícius pelo apoio incondicional durante esse período de doutoramento. E à todos e todas que direta ou indiretamente contribuíram para a realização desta tese.

POESIA: “words set to music” (Dante via Pound), “uma viagem ao desconhecido” (Maiakóvski), “cernes e medulas” (Ezra Pound), “a fala do infalável” (Goethe), “linguagem voltada para a sua própria materialidade” (Jakobson), “permanente hesitação entre som e sentido” (Paul Valery), “fundação do ser mediante a palavra” (Heidegger), “a religião original da humanidade” (Novalis), “as melhores palavras na melhor ordem” (Coleridge), “emoção relembrada na tranquilidade” (Wordsworth), “ciência e paixão” (Alfred de Vigny), “se faz com palavras, não com ideias” (Mallarmé), “música que se faz com ideias” (Ricardo Reis/Fernando Pessoa), “um fingimento deveras” (Fernando Pessoa), “criticismo of life” (Mathew Arnold), “palavra-coisa” (Sartre), “linguagem em estado de pureza selvagem” (Octavio Paz), “poetry is to inspire” (Bob Dylan), “design de linguagem” (Décio Pignatari), “lo impossible hecho possible” (Garcia Lorca), “aquilo que se perde na tradução (Robert Frost), “a liberdade da minha linguagem” (Paulo Leminski)...

“Limites ao Léu”, Paulo Leminski

RESUMO

Nas últimas décadas observa-se um crescente uso de sistemas de tradução automática para a tradução rápida e imediata de fragmentos de diversos documentos. Ao mesmo tempo, observa-se um frequente uso desses sistemas em documentos multimodais, originalmente criados de forma a relacionar semanticamente elementos linguísticos e visuais. Como as máquinas de tradução não são desenvolvidas com o propósito de reconhecer tais relações intersemióticas, mas apenas de reconhecer determinadas unidades linguísticas, pressupõe-se que parte do significado originalmente criado entre o componente linguístico e visual se modifica com a tradução. Um estudo piloto, produzido como parte do estágio metodológico deste trabalho, aponta para desvios léxico-semânticos de legendas que descrevem uma imagem ou parte dela, gerando uma nova configuração semântica, chamada de incompatibilidade intersemiótica. Por meio da interface teórico-metodológica da multimodalidade, mais especificamente do modelo de textura intersemiótica, e da linguística computacional, mais especificamente de tipologia de erro de tradução automática, esta tese tem como objetivo identificar relações semânticas entre texto e imagem em documentos multimodais traduzidos automaticamente para a avaliação de tradução automática. Os procedimentos metodológicos para se alcançar o objetivo de pesquisa são de natureza exploratória e qualitativa. A amostragem é constituída de 30 documentos compostos por artigos de páginas web, manuais, infográficos e artigos acadêmicos em inglês e suas respectivas traduções automáticas realizadas pela ferramenta Google Tradutor para o português. Para a captura das páginas web utilizou-se a ferramenta *Fireshot*, e para visualização e anotação manual escolheu-se as ferramentas *Preview* e *Sketch*. A análise revela um número proporcionalmente maior de incompatibilidades intersemióticas encontradas nas páginas web do que nos outros tipos de documentos. Encontrou-se os mesmo tipos de incompatibilidade intersemiótica que emergem no estudo piloto, quais sejam, de relação intersemiótica temporal e relação intersemiótica de ambiguidade. Porém identificou-se e categorizou-se outros níveis dessas relações, tais como as relações intersemióticas de correspondência e relações intersemióticas de contingência. Os achados também revelam um padrão de incompatibilidades intersemióticas advindas de problemas de palavras desconhecidas e palavras incorretas (VILAR et al., 2006). Ao final, os achados da análise demonstram a importância do modelo de textura intersemiótica proposto por Liu e O'Halloran (2009) para se elaborar uma base metodológica para novas tipologias de erros de tradução automática que identifiquem e categorizem relações de texto-imagem para a avaliação manual de tradução automática.

Palavras-chave: Tradução automática; Multimodalidade; Avaliação de Tradução Automática; Textura Intersemiótica; Incompatibilidade Intersemiótica.

ABSTRACT

In the past decades there has been a growth in the use of automatic translation systems for the fast and immediate translation of diverse document fragments. At the same time, there has been a frequent use of such systems in multimodal documents, originally created to relate semantically linguistic and visual elements. As machine translation systems are not developed with the purpose of recognizing such intersemiotic relations - but to only recognize certain linguistic unities - this study assumes that part of the meaning originally created across the linguistic and visual components changes with the translation. A pilot study produced as part of the methodological stage of this work shows that lexico-semantic divergences of automatically translated captions which describes an image or part of it generates new semantic configuration, namely intersemiotic mismatch. By means of the theoretical and methodological interface of multimodality, more specifically the intersemiotic texture model proposed by Liu e O'Halloran (2009), and of computational linguistics, more specifically the typology of machine translation errors (VILAR et al., 2006), this dissertation has the objective of identifying semantic relations between text and image in multimodal documents automatically translated for automatic translation evaluation. The methodological procedures are based on exploratory and qualitative nature. The sampling consists of 30 documents made of webpages, manuals, infographics, and academic papers in English and their respective automatic translation output generated by means of Google Translator into Portuguese. To capture webpages this investigation employs the tool Fireshot, and to visualize and manually annotate the intersemiotic relations, it employed the tools Preview and Skitch. The analysis reveals a proportionally greater number of intersemiotic mismatches found on the webpages than on the other documents. The study found the same types of intersemiotic mismatches that emerged in the pilot study, though in more specific levels. Among the most relevant findings of these new levels of mismatches are the intersemiotic contingency relations and the intersemiotic relations of correspondence. The findings also have shown a profile of intersemiotic mismatches related to unknown words and incorrect words (VILAR et al., 2006). At the end, the findings of the analysis show the importance of the Intersemiotic texture model proposed by Liu and O'Halloran (2009) to elaborate a methodological ground for new typology of machine translation errors that informs the identification and categorization of text-image relations for manually evaluate automatic translation.

Keywords: Machine Translation; Multimodality; Machine Translation Evaluation; Intersemiotic Texture; Intersemiotic mismatch.

LISTA DE FIGURAS

Figura 1 - Exemplo de perspectiva de avaliação <i>black box</i> de um sistema de TA	74
Figura 2 - Classificação de erros de TA.....	75
Figura 3 - Exemplos de divergências de tradução	82
Figura 4 - Categorias de textura intersemiótica propostas por Liu e O'Halloran (2009)	87
Figura 5 - Exemplo de estrutura paralela intersemiótica.....	89
Figura 6 - Exemplo de polissemia intersemiótica	90
Figura 7 - Interface teórico-analítica	95
Figura 8 - Interface interdisciplinar	101
Figura 9 - Imagens capturadas do <i>site Wikipédia</i> em inglês e sua tradução do Google Tradutor para o português.....	108
Figura 10 - Imagens capturadas do <i>site Wikipédia</i> em inglês e sua tradução do Google Tradutor para o português.....	110
Figura 11 - Imagens capturadas do <i>site BuzzFeed</i> em inglês e sua tradução do Google Tradutor para o português.....	114
Figura 12 - Exemplo 1 de ambiguidade intersemiótica.....	128
Figura 13 - Exemplo 2 de ambiguidade intersemiótica.....	130
Figura 14 - Exemplo 1 de incompatibilidade de relação intersemiótica de tempo.....	131
Figura 15 - Exemplo de incompatibilidade intersemiótica com palavras desconhecidas	134
Figura 16 - Incompatibilidade intersemiótica em artigo da <i>BBC</i>	136
Figura 17 - Incompatibilidade intersemiótica em tradução automática	137
Figura 18 - Exemplo de incompatibilidade intersemiótica em página da <i>National Geographic Kids</i>	139
Figura 19 - Infográfico da empresa <i>Brandyourself</i> com incompatibilidade intersemiótica	141
Figura 20 - Incompatibilidade intersemiótica em artigo acadêmico.....	144
Figura 21 - Proposta de incompatibilidade intersemiótica na interface de investigação	149

LISTA DE TABELAS

Tabela 1 - Linguística computacional na revista <i>Ciência da Informação</i>	51
Tabela 2 – Linguística computacional na revista <i>Perspectivas em Ciência da Informação</i>	54
Tabela 3 - Linguística computacional na revista <i>Incid: Ciência da Informação,</i> <i>Informação, Documentação</i>	54
Tabela 4 - Linguística computacional na revista <i>Informação e Sociedade</i>	56
Tabela 5 - Camadas principais da estrutura conceitual de Gênero e Multimodalidade...	67
Tabela 6 - Classificação de problemas morfológicos e léxicos do TrAva	79
Tabela 7 - Classificação de problemas sintáticos do TrAva	80
Tabela 8 – Relações intersemióticas lógicas.....	92
Tabela 9 - Informações gerais sobre os documentos analisados	120

LISTA DE GRÁFICOS

Gráfico 1 - Perfil dos tipos de documentos multimodais	122
Gráfico 2 - Perfil de documentos traduzidos total e parcialmente	124
Gráfico 3 - Perfil de tipos de conteúdo dos documentos analisados.....	125

LISTA DE ABREVIATURAS E SIGLAS

ALPAC – *Automatic Language Processing Advisory Committee*
API – *Application programming interface*
ATA – Avaliação de tradução automática
CI – Ciência da Informação
ET-ADGV – Especificação Técnica para Aquisição de Dados Geoespaciais Vetoriais
LC – Linguística computacional
MIT – *Massachusetts Institute of Technology*
OC – Organização do conhecimento
OCR – *Optical character recognition*
OI – Organização da informação
PLN – Processamento de linguagem natural
RST – *Rhetorical structure theory*
SOC – Sistemas de Organização do Conhecimento
TA – Tradução automática
TABD – Tradução automática baseada em dicionário
TABR – Tradução automática baseada em regras
TABE – Tradução automática baseada em exemplos
TAE – Tradução automática estatística
TAO – Tradução automática online
TI – Tecnologia da informação

SUMÁRIO

1. Introdução	14
1.1 Delimitação do problema	14
1.2 Objetivo geral	19
1.3 Justificativa.....	20
1.4 Organização do trabalho	21
2. Marco Conceitual: Contextualização	24
2.1 Conceituando a informação	25
2.2 Revisão histórica sobre a tradução automática.....	29
2.3 Abordagens de tradução automática	34
2.4 Avaliação de tradução automática	40
2.5 Ciência da Informação e Linguística Computacional	46
2.6 Multimodalidade	62
2.7 Multimodalidade e Ciência da Informação.....	68
2.8 Considerações Finais	72
3. Marco Teórico.....	73
3.1 Categorias de erro de tradução automática	73
3.2 Relações texto-imagem	85
3.3 Considerações finais	95
4. Métodos	97
4.1 Procedimentos metodológicos	98
4.1.1 Instrumentos e materiais	102
4.2 Estudo piloto.....	104
4.3 Resultados	115
5. Análise da base de dados	119
5.1 Análise geral	120
5.2 <i>Webpages</i>	126
5.3 Infográficos	139
5.4 Manuais	142
5.5 Artigos científicos	143
6. Considerações finais.....	146
6.1 Limitações da pesquisa	150
6.2 Sugestões e recomendações	151

7. Referências bibliográficas.....	153
ANEXO A - Parecer do co-orientador estrangeiro	161
ANEXO B - Autorização para uso do Google Tradutor.....	162
APÊNDICE A - Cronograma de atividades.....	164
APÊNDICE B - Tabela de classificação dos documentos multimodais	167

1. Introdução

O presente capítulo apresenta a delimitação do problema, o objetivo, a justificativa, e a organização do trabalho. O tema do presente estudo insere-se na linha de pesquisa da Organização da Informação, especificamente nas áreas da Multimodalidade e da Linguística Computacional, áreas que têm gradualmente se estabelecido na Ciência da Informação e com a qual pretende-se contribuir.

1.1 Delimitação do problema

Desde o século dezessete já se discutia a ideia de se mecanizar o processo tradutório. Mas foi com o advento dos computadores no período do pós-guerra, que novas disciplinas se formaram em torno do desenvolvimento desse conceito (HUTCHINS, 1986, 2010). Assim surgiram, por exemplo, áreas como a Ciência da Informação (CI) e mais especificamente a Linguística Computacional (LC), ambas interessadas na investigação de problemas pontuais de tradução automática, envolvendo ambiguidade lexical, alinhamento e análise semântica.

As temáticas similares entre a LC e a CI como parte integrante dos estudos sobre o fenômeno da informação foram levantadas no artigo seminal de Boroko (1968) logo

após a mudança no nome da *American Documentation Institute* para *American Society for Information Science*. Borko (ibid.) discute primeiramente o significado de ciência da informação. Para o autor (1968, p.1, tradução nossa) a CI é a disciplina que investiga as propriedades e comportamentos da informação, as forças que governam o fluxo de informação, e o meio de processamento da informação para acessibilidade e usabilidade ótimas¹.

Além disso, somam-se à natureza interdisciplinar da CI os aspectos “puros” e “aplicados” (BORKO, 1968, p.3), onde até então, grande parte das pesquisas da área focaram nos primeiros. Independente dos aspectos levados em conta pelo cientista da informação, Borko (1968, p. 3-4) assinala nove grandes categorias que emergem das principais pesquisas levantadas na área até então, a saber: a) usos e necessidades da informação; b) cópia e criação de documento; c) análise linguística; d) tradução; e) *abstracting*, classificação, codificação e indexação; f) design de sistema; g) análise e avaliação; h) reconhecimento de padrão; e i) sistemas adaptativos.

No que se refere à categoria tradução, pode-se identificar as subcategorias de tradução automática e auxílios à tradução. Contudo, desde a publicação deste artigo nota-se uma escassa produção acerca dos problemas gerados pela tradução automática,

¹ Entende-se aqui que os estudos sobre acessibilidade e usabilidade desenvolveram-se substancialmente desde a publicação deste artigo. É notável a complexidade de categorias e definições para o termo ‘ótimo’ (do inglês *optimum*) dentro das respectivas áreas. Contudo, subentende-se que o autor se dirija ao termo de uma forma mais geral no sentido de acesso e uso rápido e eficiente.

sobretudo dentro do contexto da ciência da informação no Brasil, isto é de uma perspectiva de ciência social aplicada aos estudos de tradução automática.

Nessa direção, nota-se desde a publicação do artigo de Borko (1968), a crescente produção em algumas dessas categorias e na interface entre elas, naturalmente, contribuindo para o estabelecimento da CI como disciplina independente e de caráter interdisciplinar. Um dessas áreas é a organização do conhecimento, definida por Hjørland (2008b) a partir de como a informação é organizada socialmente e de como a realidade é organizada (dentro de um escopo mais abrangente²); e (dentro de um escopo mais específico) argumenta que a organização da informação na CI envolve atividades tais como descrição, indexação, e classificação de documentos.

Ambas perspectivas do conceito de organização do conhecimento podem funcionar juntas quando se pensa em Sistemas de Organização do Conhecimento (doravante SOC³) (HJORLAND, 2008), definidos como um termo mais geral para “as ferramentas que apresentam a interpretação de estruturas do conhecimento”. De acordo com Hjørland (ibid.), o mesmo termo corresponde às ferramentas semânticas tais como os thesauri, ontologias, dicionários que produzem “informações semânticas”, isto é, informações sobre o significado das palavras e outros símbolos referentes às

² Os termos do original em inglês *Broad sense* e *narrow sense* são desenvolvidos em Hjørland (2008).

³ Do inglês *Knowledge Organization Systems (KOS)*.

relações entre símbolos e conceitos.

A partir do conceito de ferramenta semântica de Hjørland (2007), nota-se uma escassez de estudos que investiguem a relação dos significados das palavras com outros sistemas de significação (como por exemplo o visual) em documentos traduzidos automaticamente. Tais fenômenos podem ocorrer naturalmente com o crescente uso de tradutores online em documentos onde língua e imagem são semanticamente interligados para se construir significado, tais como fotografias e suas legendas em artigos de página *web*.

Websites de conteúdo variado são documentos multimodais⁴ frequentemente requisitados para a tradução online e gratuita, frequentemente realizada pelo *Google Translate*⁵. Mas o que é que dessa “tradução” pode ser considerada informação deve ser no mínimo parte da “combinação de diferentes modos de informação” (BATEMAN, 2008), gerada na forma apresentada ao usuário. Em outras palavras, a forma como aspectos visuais e verbais são combinados podem revelar significados em potencial para serem traduzidos.

A abordagem comunicativa e sócio-semiótica interdisciplinar da multimodalidade (KRESS; VAN LEEUWEN, 1996, 2001) tem nas últimas décadas

⁴ De acordo com Bateman (2008, p. 1, tradução nossa) documento multimodal é um tipo de artefato multimodal contendo uma variedade de modos baseados no visual e que organizam-se simultaneamente para preencher uma coleção orquestrada de objetivos comunicativos interligados entre si.

⁵ Disponível em <https://translate.google.com/>. Acesso em: 19 jul. 2016.

desenvolvido um número crescente de trabalhos sobre práticas comunicativas que utilizam recursos visuais, escritos, auditivos, e espaciais (chamados de “modos”) utilizados para compor mensagens.

Embora existam descrições e categorias de análise na relação texto-imagem, há ainda escassa pesquisa que investiga tal relação com resultados de traduções automáticas. E o inverso é também observado na área da linguística computacional, onde há uma escassez na literatura, especificamente na avaliação manual de resultados de tradução automática (área da Avaliação de Tradução Automática) que investigue essa relação a partir de tipologias de problemas dos resultados gerados por máquina de tradução.

A utilização do termo “erros” dentro do contexto da computação é amplamente utilizado no sentido matemático (no sentido do “cômputo”). Por essa razão é raramente conceitualizado, questionado ou teorizado quando o mesmo é relacionado a uma aplicação de tradução automática. Assim, a presente pesquisa lança mão de uma perspectiva sócio-semiótica (multimodalidade) sobre tais “erros” em um dado contexto de produção. Portanto, entende-se no presente estudo que “erros de tradução” são configurações de incompatibilidades linguísticas (lexical, semântica, e sintática) entre o texto de entrada e o texto de saída gerado por uma tradução automática em um dado contexto de produção (por exemplo, de um determinado artigo acadêmico ou *webpage*).

Tal observação se faz necessária, já que o termo “erro de tradução” é constantemente problematizado do ponto de vista da área de Letras, e mais especificamente dos estudos da tradução, áreas que contribuem para a formação da perspectiva investigativa juntamente com a computacional dentro do âmbito da organização do conhecimento, e da ciência da informação.

O problema de pesquisa delimitado é também calcado nos achados do estudo piloto (vide seção 4.2), que apontam para determinados tipos de relação semântica entre texto e imagem encontrados a partir de resultados gerados pelo serviço online do Google Tradutor.

Considerando esses aspectos, formula-se as seguintes perguntas de pesquisa:

- **Como a abordagem da multimodalidade contribui para a avaliação de resultados de tradução automática em documentos multimodais?**
- **Como se dão as relações texto-imagem em documentos multimodais a partir dos resultados da tradução automática?**

1.2 Objetivo geral

A presente investigação parte do pressuposto de que há relações de significado entre elementos linguísticos e visuais possíveis de ser investigados no contexto onde

uma tradução automática foi gerada.

Assim, o objetivo geral da pesquisa é identificar relações semânticas entre texto e imagem em documentos multimodais traduzidos automaticamente para a avaliação de tradução automática.

1.3 Justificativa

O presente estudo não tem como propósito fazer uso da multimodalidade para solucionar um problema de tradução automática, tampouco busca resolver um problema da subárea de avaliação de tradução automática. Esta investigação se propõe a expandir o olhar do pesquisador para a informação traduzida automaticamente em textos multimodais. Em outras palavras, busca examinar como a abordagem da multimodalidade pode contribuir para a identificação de relações texto-imagem geradas a partir de traduções automáticas.

Bateman (2008, p. 7) oferece um modelo para a análise sistemática de documentos multimodais, os quais “carregam significados que se utilizam de *modos* visuais, espaciais e verbais em combinação e em cooperação”. Em outras palavras, a multimodalidade pretende preencher essa perspectiva, pois tais significados não são encontrados dentro do paradigma tradicional da linguística ou do *design* separadamente.

Assim, a convergência da abordagem da multimodalidade e da linguística computacional podem contribuir para a CI, sobretudo para as áreas da Organização da Informação e do Conhecimento. O alcance dos resultados de tecnologias da tradução também se estendem na medida em que há evidentes elementos coesivos de diferentes componentes semióticos, construídos especificamente para um determinado tipo de documento multimodal. Nesse âmbito, ambas perspectivas podem contribuir para a investigação de padrões dessas relações e, futuramente, para o embasamento de modelos que auxiliem na tarefa de avaliação de tradução automática de documentos multimodais.

1.4 Organização do trabalho

Este trabalho está organizado da seguinte forma:

O *capítulo 1* fornece uma breve delimitação do problema de pesquisa apontando para a necessidade de se abordar as relações semânticas entre texto e imagem a partir da geração de tradução automática dentro do âmbito da Multimodalidade e da Linguística Computacional. Tal delineamento sugere duas perguntas de pesquisa que norteiam o objetivo geral, seguido da justificativa para a realização do trabalho.

O *capítulo 2* elabora a contextualização do marco conceitual da pesquisa. Aqui se detalha a interface investigativa compreendendo alguns conceitos de informação da CI,

como área mais abrangente; elabora um levantamento da literatura sobre linguística computacional disponível dentro da CI; levanta as primeiras abordagens sobre tradução automática; e revisa-se alguns conceitos da literatura da área da multimodalidade e sua aplicação na CI no Brasil.

O *capítulo 3*, descreve o Marco Teórico que aprofunda os estudos nas duas interfaces teórico-analíticas, quais sejam a multimodalidade e a linguística computacional para as relações intersemióticas (por meio do conceito de textura intersemiótica), e da perspectiva de classificação de erros de tradução automática por meio da perspectiva *black box* (a partir dos resultados apresentados pela máquina, sem levar em conta a engenharia do sistema) de avaliação de tradução automática a fim de informar o processo de análise.

O *capítulo 4* traça os procedimentos metodológicos utilizados para se atingir o objetivo de pesquisa, discute inicialmente a criação do desenho metodológico mais apropriado, e elabora os passos para a execução do estudo piloto. Os achados do estudo piloto redefinem alguns aspectos metodológicos que são levados em consideração na composição da base de dados (amostra, no sentido mais amplo) para posterior análise.

O *capítulo 5* desenvolve a análise de documentos multimodais compostos por alguns subtipos de páginas *web*, manuais, infográficos, e artigos acadêmicos. O capítulo inicia com um breve panorama quantitativo sobre a amostra e na sequência discute o

processo de identificação e classificação das relações texto-imagem da referida base de dados .

O *capítulo 6* desenvolve as considerações finais, retomando as perguntas de pesquisa e objetivo proposto, discute o percurso metodológico, passando pelo estudo piloto e suas redefinições para a análise da base de dados. Apresenta também as limitações encontradas durante a investigação e sugere recomendações para possíveis pesquisas futuras.

2. Marco Conceitual: Contextualização

Este capítulo tem como objetivo descrever o campo de investigação no qual a pesquisa proposta se insere. Sobretudo, este capítulo busca evidenciar o amplo caráter interdisciplinar do presente problema de pesquisa, situado, a grosso modo, nas áreas da organização da informação, multimodalidade e linguística computacional. Para tanto, a primeira subseção conceitua a informação a partir de autores da ciência da informação (CI) e da linguística computacional (LC); a segunda subseção localiza a tradução automática (TA) dentro de um conciso traçado histórico; a terceira subseção conceitua as principais abordagens de tradução automática, com uma parte especialmente dedicada a tratar do serviço de tradução online *Google Translate*; a subseção seguinte conceitua a área de avaliação de tradução automática (ATA), com um enfoque na categorização manual dos chamados “erros” de tradução automática; a quinta subseção faz um levantamento do tema da linguística computacional em periódicos publicados na área da ciência da informação no Brasil; a sexta subseção conceitua a perspectiva da multimodalidade que delinea a pesquisa; por fim, a sétima subseção apresenta alguns estudos na interface da multimodalidade e ciência da informação no Brasil.

2.1 Conceituando a informação

Há algumas décadas, a área de ciência da informação (CI) foi criada com o intuito de "investigar as propriedades e comportamento da informação, as forças governantes do fluxo de informação, e os meios de processamento da informação para a acessibilidade e usabilidade ótimas" (BORKO, 1968, p.3). Desde então nota-se uma intensa e diversa busca pela definição do conceito de informação.

Para Jason Farradane (FARRADANE, 1976, p. 13), químico inglês dedicado à documentação, integrante do *Classification Research Group (CRG)*:

A informação deve ser definida como qualquer forma física de representação, ou substituto do conhecimento, ou um pensamento particular, usado na comunicação. [Sua] definição de informação como um substituto escrito ou falado do conhecimento é mais explícita que o tratamento da informação como um conceito: pode-se tratar um registro escrito permanente como um ponto de partida invariante do qual se pode derivar, na medida do possível, de maneira experimental, as evidências de sua conversão [do registro do conhecimento] do pensamento do emissor, ou para o pensamento e as consequências deste no indivíduo que o percebe e em outros.

A primeira definição foi proposta por Farradane, em 1976, e analisada por Belkin (1978) em artigo que o estudioso aborda várias definições de informação. Belkin (1978, p.78) afirma que o conceito proposto por Farradane (1976), corresponde à noção de informação do senso comum e é também utilizado quando se dispõe de poucas palavras para descrever o que é informação, ou, quando é apresentado informalmente um

conceito de informação antes que outro possa ser mais rigorosamente definido.

Belkin (1978) explicita que o conceito de Farradane (1976) não se limita àquele do senso comum, pois Farradane (1976) aborda o campo de estudo da ciência da informação no qual insere os seres humanos, seus processos de pensamento e comportamento em todos os eventos relacionados à comunicação, isto é, os sujeitos envolvidos no processo comunicativo. Contudo, Belkin (1978) salienta que nesta definição não é contemplado o efeito da informação no destinatário, fato apontado pelo próprio Farradane (FARRADANE, 1979), e proposto de ser revisto.

A segunda definição corresponde àquela publicada em artigo intitulado *The nature of information* (FARRADANE, 1979), que revisita o conceito proposto em 1976 e incorpora as observações feitas por Belkin (1978). Ressalta-se que na seção em que conceitua a informação, no artigo citado, Farradane também a caracteriza como “a declaração de um emissor, ou a solicitação feita a um sistema de recuperação da informação” (FARRADANE, 1979, p.13), sendo estas, para o estudioso, do mesmo tipo, mas que requerem análise de forma distinta. Farradane (1979) frisa que outros estágios de comunicação que, não estão explicitamente expressos em um sistema de recuperação da informação, são mentais em sua natureza, e não podem ser diretamente examinados⁶.

⁶ As passagens onde constam levantamento dos trabalhos de Farradane e Belkin são parte de artigo não-publicado de autoria do presente candidato à doutoramento e de Alessandra Rodrigues.

Brookes (BROOKES, 1980) faz a distinção entre o termo informação e conhecimento. Para o autor (ibid.) conhecimento é como uma estrutura de conceitos ligados por suas relações e informação como uma pequena parte de tal estrutura. A estrutura do conhecimento pode ser subjetiva ou objetiva (BROOKES, 1980, p.10).

Ainda nessa distinção, Capurro e Hjørland (2003), com base em Bounoux (1993), consideram que conhecimento e informação são inversamente relacionados, uma vez que o primeiro diz respeito à previsibilidade e redundância, enquanto o segundo, ao novo e ao imprevisto (ibid., p.173); então, para os autores, informar significa "selecionar" e "avaliar". Contudo, Capurro e Hjørland (2003) problematizam a busca pela definição da informação isoladamente, alegando uma confusão⁷ dentro da CI sobre o conceito de informação, enquanto notadamente área interdisciplinar.

Em seu trabalho, Capurro e Hjørland (2003) sinalizam que o padrão de divergência sobre a discussão de informação reside não apenas em olhar a informação como objeto, mas, em uma perspectiva mais profunda, na natureza dos mecanismos de processamento de informação, os seletores e intérpretes - ideia do mecanismo de liberação que advém de Karpatschhof (2000 p. 131-132 apud CAPURRO; HJORLAND, 2003, p.176).

Já na fronteira da Ciência da Informação e da Tecnologia da Informação, observa-

⁷ Tal confusão na Ci diz respeito a certa fluidez conceitual, à qual ela se presta, devido, possivelmente, às suas próprias características interdisciplinares.

se a investigação de alguns autores que trabalham o conceito de informação dentro do âmbito da tradução automática. W. John Hutchins, originalmente bibliotecário na universidade de Sheffield, Inglaterra, cruzou conhecimentos linguísticos para a computação de processos bibliotecários. Embora o trabalho do bibliotecário e acadêmico não seja conhecido dentro da CI, o mesmo é notadamente reconhecido na interface dos estudos da tradução e da ciência da computação.

Dentro do escopo histórico da tradução automática da década de 30 aos dias atuais, Hutchins (2006) classifica o uso de serviços de tradução automática online como uma “tradução grosseira e imediata” com fins de informação - algo oferecido por sistemas *mainframe* nos anos 60, e geralmente ignorado anos depois (ibid, p. 17). Em outras palavras, identifica-se nessa classificação de Hutchins (2006) a questão cognitiva da informação em função das limitações da máquina de tradução quando os serviços de tradução online são utilizados.

Como o objeto de estudo proposto circunda a informação gerada a partir de um sistema de tradução automática, faz-se necessário subsidiar essa informação traduzida automaticamente com uma discussão histórica. Assim, a subseção a seguir trata de revisar alguns aspectos históricos sobre a TA.

2.2 Revisão histórica sobre a tradução automática

Conforme mencionado na introdução deste estudo, a ideia de se traduzir uma língua "mecanicamente" remete ao século dezessete quando filósofos discutiam conceitos de "línguas universais" e dicionários "mecânicos", conforme apontado por Hutchins em recentes revisões históricas acerca da tradução automática (TA) (HUTCHINS, 1986, 2010). Mas a ideia comumente compartilhada hoje em dia da TA enquanto uma ferramenta computacional que fornece uma tradução inteiramente automática de uma língua para outra foi colocada em prática apenas a partir do advento do computador na década de 40, mais precisamente a partir do Memorando Weaver (HUTCHINS, 1997).

Conforme as revisões de Hutchins destacam, o Memorando Weaver é um documento escrito por Warren Weaver, convidando entidades de diversos ambientes políticos, financeiros e científicos que poderiam investir em desenvolvimentos na tradução mecanizada ou TA. É dentro do contexto do pós-guerra que as questões de máquinas de tradução são discutidas dentro de um âmbito acadêmico. E a partir daí inicia-se uma área interdisciplinar formada por cientistas da computação, linguistas e matemáticos, entre outros.

Por mais complexo que o tópico de tradução automática possa ser, (e apesar de seu desempenho rudimentar e relativo sucesso nessa época) sua criação tornou-se o

ponto inicial de um campo emergente de estudos conhecido como Processamento de Língua Natural (PLN) ou Linguística Computacional (cada uma refere-se ao mesmo problema a partir de uma perspectiva mais computacional ou linguística). Em relação ao tópico, Nunes (2008, p.4) ressalta que a PLN “começou do fim”, isto é, um sistema sofisticado como a TA, que requer uma combinação de soluções em diversos níveis linguísticos e computacionais, se destaca inicialmente na área ao invés de sistemas menos complexos como sumarização ou extração da informação. Em seu artigo, Nunes (ibid.) fornece uma breve introdução da PLN e inclui a TA.

Em geral os dois autores abordam o tópico da tradução automática, porém a presença de Hutchins é notória na comunidade da linguística computacional por propor uma análise mais “humana” (em oposição à computacional), “de fora da máquina” de tradução automática, e mais especificamente a partir de um ponto de vista mais histórico.

Já Nunes, também explora o tópico da TA, porém como parte de um raciocínio de problema-solução de PLN, utilizando-se de métodos e ferramentas para o desenvolvimento de solução para problemas de TA envolvendo o português brasileiro. Apesar das duas abordagens, histórica ou computacional, ambas as visões levam em consideração os vários sistemas nos quais TAs se baseiam.

Hutchins (2010, 2015) elabora uma revisão concisa dos sistemas e projetos mais

populares de pesquisa sobre TA dos anos mais remotos ao período dos serviços de TA na internet.⁸ O autor divide a história da TA em pesquisa "pioneira" e primeiros sistemas operacionais dos anos 40 aos 60, à hegemonia dos sistemas baseados em regras dos 60 aos 80, à emergência dos sistemas baseados em corpus em ambiente recente de internet.

Hutchins (2010, p.1) inicia sua revisão a partir do marco de duas patentes de "dicionários mecânicos" (embora desconhecidas para Warren Weaver e Andrew Booth⁹) que apresentam uma proposta para lidar com problemas de ambiguidade baseados em criptografia, estatística, teoria da informação, lógica e universais da linguagem. As universidades de MIT e Georgetown University, lideradas por acadêmicos como Bar-Hillel¹⁰ e León Dostert,¹¹ foram respectivamente pioneiras em pesquisa de TA.

Hutchins (2010) destaca como primeiros eventos acadêmicos sobre TA o início das parcerias entre setores acadêmicos e privados e a primeira demonstração de tradução automática do russo para o inglês. De acordo com o autor (ibid., p. 2), tais eventos podem ser vistos como os primeiros marcos na pesquisa de TA situada dentro

⁸ Para mais informações sobre os registros de Hutchins sobre o desenvolvimento de TA vide Hutchins (1986, 1988, 1994, 1999).

⁹ Pioneiros no desenvolvimento de tradução automática na segunda metade da década de 40.

¹⁰ Filósofo, matemático e linguista israelense. Para maiores informações vide Hutchins (HUTCHINS, 2000).

¹¹ Professor de língua francesa responsável por lançar o primeiro experimento com TA em colaboração com a IBM.

do contexto político da Guerra Fria.

A mais concisa história da TA elaborada por Hutchins (2015) resume esse período e enfatiza a pesquisa em TA que se iniciou a partir dos relatórios da ALPAC (*Automatic Language Processing Advisory Committee*) (1966), quando o governo norte-americano cancelou o financiamento alegando que os resultados e utilidade de tradução integralmente automatizada eram "inviáveis para cumprir as expectativas das pessoas", mas apontando desenvolvimentos mais úteis das traduções assistidas por computador (HUTCHINS, 2015, p.4).

Embora cada trabalho destaque diferentes áreas da tecnologia da tradução, isto é da tradução automática e da tradução assistida por computador, (ou seja tradução realizada com a assistência de ferramentas de computação tais como memórias de tradução, base de dados bilíngue e sistemas de gerenciamento de tradução) ambas as perspectivas se complementam. Conforme apontado por Hutchins (2015, p.3), "reconhece-se em geral que a falha da tradução automática nos anos 60 [geralmente referindo-se ao relatório "infame" da ALPAC em 1966] levou à emergência da tradução assistida por computador".

Apesar da redução no volume de investimento para se desenvolver a TA, na década seguinte ao relatório da ALPAC ainda havia alguns grupos desenvolvendo sistemas de TA. Conforme Hutchins (2010, p.6), a primeira geração de pesquisadores

em TA priorizava a “tradução direta”, e nos anos seguintes ao relatório da ALPAC, trabalhavam na tradução automática “indireta tais como interlândia e sistemas baseados em transferência. De acordo com o autor (ibid., p. 6), dentre os sistemas que ganharam maior destaque no fim dos anos 60 e 70, está o projeto TAUM (*Traduction Automatique de l’Université de Montréal*), que elaborou um formalismo para manipular *strings*¹², árvores linguísticas, e o sistema METEO para “traduzir boletins de previsão do tempo”.

Hutchins (2010, p.7-8) destaca o período de 1976 a 1989 em termos de relevância comercial e acadêmica. Empresas de computação no Japão começaram a desenvolver *softwares* para tradução auxiliada por computador.

Tal iniciativa levou a um crescimento de opções a um *workbench* de tradução, isto é, o profissional da tradução tinha controle sobre os processos de tradução por meio da popularização dos computadores. Ao mesmo tempo, pela influência das teorias linguísticas sobre formalização de regras lexicais e gramaticais, projetos como o EUROTRA, um sistema de transferência multilíngua que objetivava a tradução de todas as línguas da comunidade europeia, começaram a ganhar nova dimensão na pesquisa de TA (HUTCHINS, 2010, p.9).

Baseado em referências da *Annual Machine Translation Archives* (de 1991 a

¹² sequência de caracteres.

2006), Hutchins (2010, p.11) aponta para uma mudança de paradigma no final dos anos 80, das abordagens baseadas em regras para a aplicação de métodos baseados em corpus em TA . O autor justifica que esse movimento deve-se à precisão dos modelos baseados em estatística demonstrado pelos modelos de reconhecimento de voz da IBM, e que embasa sua informação probabilística em *corpora* bilíngues ou multilíngues. O artigo termina com uma consideração sobre a TA na internet. Hutchins (2010, p. 17) cita exemplos de traduções online de *webpages* e *emails* utilizando *Systran*¹³ e outros sistemas como *Altavista* e *Google*.

A partir da discussão histórica da tradução automática, nota-se a relevância de se contextualizar os diferentes tipos de abordagem de máquinas de tradução. Assim, a próxima subseção propõe a discussão dos principais tipos de abordagem de sistemas de TA afim de delimitar parte da presente investigação.

2.3 Abordagens de tradução automática

Conforme visto na revisão histórica de TA, muitos são os paradigmas desenvolvidos (e em desenvolvimento) sobre a tradução automática. Contudo, alguns deles são frequentemente aplicados e discutidos pela literatura disponível. Assim esta subseção objetiva a revisão de algumas referências que discutem as abordagens de TA

¹³ Disponível em: ([HUTCHINS, 1986, 1988, 1994, 1999](#)). Acesso em: 20 jul. 2016.

seguintes.

- **Tradução automática baseada em dicionário (TABD)**

De acordo com Hutchins, (1986, 2001, 2010) os primeiros *insights* sobre tradução automática que podem traduzir de uma língua para outra vieram com o dicionário mecânico multilíngue, especialmente nos anos 40 e 50. Esse tipo de tradução automática é baseada em dicionários bilíngues de forma legível para a máquina. Para Tripathi e Sarkhel (2010, p.389), em seu panorama de abordagens para a TA, este método é ainda utilizado com sucesso na tradução automatizada de grupos frasais, mas não de sentenças. Embora a abordagem seja um método mais básico (geralmente baseado em análise morfológica ou técnicas de lematização¹⁴), ainda é utilizado em combinação com outras abordagens de TA tais como aquelas baseadas em exemplo e em regras.

- **Tradução automática baseada em regras (TABR)**

Embora a tradução baseada em dicionário tenha sido a primeira a ser implementada em anos recentes do processamento de linguagem natural, Lagarda et al.

¹⁴ processo de remoção do final inflexional de uma palavra para retornar a sua raiz chamada "lema"; por exemplo, *walk - walks, walking, walked*.

(2009) consideram a metodologia de tradução automática baseada em regra (TABR) como a primeira abordagem “clássica” aos Sistemas de Tradução Automática. Esse sentido clássico significa que até fim dos anos 80 a maior parte dos sistemas de TA utilizava metodologias baseadas em regras (que descrevem conhecimento sintático) e léxico (os quais referem-se à informação semântica, morfológica, e sintática (LAGARDA et al., 2009)).

- **Tradução automática baseada em exemplos (TABE)**

Diferentemente das abordagens baseadas em dicionário, as abordagens baseadas em exemplos podem ter diferentes visões na literatura, conforme elaborado por Hutchins (2005). Em geral, o autor (HUTCHINS, 2005, p. 68) destaca que a essência da TABE é detectar correspondência de fragmentos de uma língua-fonte (em um texto de entrada) com fragmentos da língua-fonte (numa base de dados) e a extração de fragmentos da língua-alvo (como potenciais traduções parciais).

Porém, Hutchins (ibid.) destaca que há uma confusão generalizada na tentativa de se definir os sistemas de TABE e propõe um conceito de TABE mais claro em relação a tradução automática baseada em estatística e TABR. Segundo o autor (HUTCHINS, 2005, p.70), a tentativa de definir TABE é fornecer aos pesquisadores e observadores um “arquétipo” (comparável às definições de sistemas de TABR) que distingue sistemas

baseados em transferência e sistemas baseados em interlândia, enquanto na prática poucos sistemas operacionais estão em conformidade com o arquétipo em todos os detalhes .

- **Tradução automática estatística (TAE)**

Estes sistemas de TA baseiam-se em modelos estatísticos gerados a partir da análise de *corpora* bilíngues paralelos (isto é, uma coletânea de textos autênticos disponíveis eletronicamente em uma língua alinhada com sua tradução direta, ambas preparadas para posterior processamento). Para o autor Phillip Koehn, professor de tradução automática da Universidade de Edinburgh, este sistema tornou-se um dos mais utilizados e pesquisados nos últimos anos principalmente pelos serviços gratuitos de tradução (baseadas em sistemas estatísticos).

Sua principal obra de referência, *Statistical Machine Translation* (KOEHN, 2010) descreve os principais fundamentos, tais como recursos de *corpora* e teoria da probabilidade, métodos principais tais como modelos linguísticos e modelos baseados em palavras, além de tópicos mais avançados, métodos posteriores, e combinações de sistemas.

Como objeto de pesquisa e representante notório dos sistemas de tradução automática baseado em métodos estatísticos, torna-se crucial contextualizar o Google

Tradutor nesta subseção.

O Google Tradutor, ou *Google Translate* é um dos mais notáveis sistemas de tradução automática baseada em método estatístico da atualidade. Seu serviço fornece a tradução completamente automatizada de textos, imagens (decodificadas por meio de *OCR*¹⁵), *sites* e vídeos em tempo real de uma língua para outra. Além disso, o serviço oferece tradução automática para mais de 100 línguas e para mais de 200 milhões de usuários por dia¹⁶.

Um dos critérios que contribui para seu amplo acesso é a integração do sistema de TA em navegadores, tais como *Google Chrome e Firefox*. Para esse último tipo de navegador há diversos tipos de extensões para se utilizar os serviços de tradução automática do Google a partir de um clique com o botão direito, e para o navegador homônimo o serviço já é integrado automaticamente. Tais recursos favorecem o frequente uso do sistema de tradução automática *online* em *websites*¹⁷.

Além da integração em navegadores, o serviço de tradução automática do Google disponibiliza um *API (Application programming interface)* para desenvolvedores que pretendem utilizar o sistema em seus *softwares* ou *websites*. A empresa

¹⁵ Do inglês, *Optical Character Recognition*, diz respeito à forma mecanizada de se converter a imagem de caracteres escritos à mão ou impressos em texto codificado para a leitura da máquina.

¹⁶ Números indicados de acordo com o *site* do *Google Translate*, www.googletranslate.com. Acessado em 10 de novembro de 2016.

¹⁷ O presente estudo emprega esse recurso para a tradução automática das *webpages* coletadas para análise.

disponibiliza esse serviço por meio de pagamento. Contudo, a *Google* oferece o serviço do Google Tradutor API sem custos para fins de pesquisa, por meio do programa de pesquisa *Research at Google*¹⁸.

A presente investigação tem conhecimento do *API* e do programa de pesquisa oferecido pelo Google. Contudo, a ferramenta não é levada em consideração por conta do método de avaliação ser do tipo *black box* (WHITE, 2003), ou seja, leva-se em consideração a análise de erros dos resultados gerados pelo Google Tradutor, em oposição à análise de erros via *glass box*, ou seja, do exame dos mecanismos internos do sistema.

A frente do desenvolvimento do Google Tradutor desde 2007, Franz Joseph Och deu preferência à abordagem baseada em sistema estatístico, em oposição ao sistema de tradução automática baseada em regra¹⁹. O criador do modelo de tradução automática utilizado até hoje pelo Google Tradutor se vale dos critérios de custos e velocidade de processamento para uma larga escala de documentos. Baseado em *corpora* bilíngues paralelos, isto é, coletâneas de textos em uma dada língua e suas respectivas traduções contendo centenas de milhões de palavras, além de *corpora* monolíngues (textos sem tradução, utilizados como referência), contendo ainda maior

¹⁸ Para maiores informações vide <http://research.google.com/>. Acessado em 11 de novembro de 2016.

¹⁹ Até 2007 o Google Tradutor utilizava os algoritmos do Systran, baseado em regras gramaticais, e ainda utilizado pela *Microsoft*.

número de palavras, mecanismos de inteligência artificial buscam padrões por meio de cálculo estatístico e probabilístico de forma rápida da tradução mais "apropriada" (dentro do conjunto de dados que a máquina dispõe) para uma determinada entrada (OCH, 2005).

A maior parte desses textos são compostos de documentos das Nações Unidas, traduzidos entre suas 6 línguas oficiais, e a outra parte é composta por textos bilíngues disponibilizados por pesquisadores (OCH., 2005). Além do modelo estatístico proposto, o sistema de TA do Google faz uso de textos em inglês como língua intermediária entre a língua-fonte e sua tradução para a língua-alvo. Tal procedimento torna o processamento mais ágil e disponibiliza tradução automática para um maior número de línguas (BOITET et al., 2010).

Esta seção procurou revisar as principais abordagens de tradução automática. A próxima seção aprofunda a discussão em uma área específica da TA, qual seja aquela da avaliação de tradução automática (ATA), que se encarrega de avaliar os resultados e desempenho da qualidade de tais traduções.

2.4 Avaliação de tradução automática

Desde os primórdios das investigações sobre tradução automática (desde os

relatórios da ALPAC), segundo Santos et. al. (2004) a questão da qualidade das traduções desperta a atenção de estudiosos e profissionais da área. Não obstante, tornou-se importante subárea da área de processamento de linguagem natural (PLN), ou até mesmo para alguns uma área com relativa independência, (por conta de congressos especificamente de ATA) com características e problemas próprios e distintos da aplicação de TA.

John White (2003) reúne aspectos centrais da avaliação de tradução automática. O texto explora três questões centrais, quais sejam: “por que a tradução é tão central para a TA, por que a avaliação é tão difícil de se executar, e por quê deve haver diferentes tipos de avaliação para tantos tipos de usuários e usos” (WHITE, 2003, p. 211, tradução nossa). O autor (ibid.) descreve o papel da avaliação na TA com suas tradições, importância e dificuldades. Além disso, White (2003) ilustra três dos métodos principais de avaliação humana, as partes envolvidas no processo de ATA e como elas determinam seis tipos de avaliação de TA, que podem ser realizadas manualmente e/ou automaticamente, com auxílio de programas como o BLEU²⁰ (*Bilingual Evaluation Understudy*).

White (2003, p. 241) conclui que as análises de tradução automática são extremamente dependentes de aspectos “abstratos” e “intuitivos e difíceis de se

²⁰ Algoritmo criado com o propósito de avaliar automaticamente a qualidade de um texto traduzido automaticamente de uma língua para outra.

quantificar e se extrair julgamentos úteis sobre a “fidelidade” e “inteligibilidade” de resultados de tradução automática. Contudo, determinados atributos (desde aspectos linguísticos a custos operacionais) podem ser medidos dentro de determinadas variáveis passíveis de maior controle de usos e usuários.

Vilar et. al (2006) aprofundam alguns aspectos de ATA sintetizados por White (2003), e de certa forma apresentam mais compatibilidade teórica com os aspectos delimitados nesta pesquisa. Este estudo analisa erros do sistema (a partir da classificação de erros baseada em Llitjós et. al (2005). Segundo os autores (LLITJÓS et. al. 2005, p. 697), diversos métodos que medem o desempenho de um resultado de tradução automática em um dado sistema não são claros e não são possíveis quando analisados isoladamente.

O método de comparação entre o resultado da tradução gerado automaticamente e traduções consideradas “corretas”, ainda que humano, é abstrato (VILAR et al., 2006, p. 698). Os termos “correta” e “cuidado” não são definidos pelos autores (ibid.), portanto, são “carregados” semanticamente. A conceituação desses termos, acomodada em uma determinada teoria poderia contribuir para a delimitação subjetiva da análise. Pois, como afirma White há o problema da falta de verdade absoluta (*ground truth*²¹) na tradução (WHITE, 2003, p. 214) conforme o excerto abaixo:

²¹ Termo utilizado frequentemente nas ciências exatas para se referir a informações obtidas através de observação direta em oposição à inferência.

MT evaluation is harder than this. Only people who know both languages can know just by looking whether it got a translation right. And as we noted above, there is great latitude for disagreement about what constitutes “exactly right” in translation. So we cannot take full advantage of the notion of “ground truth”: the set of right answers that form a universally agreed-upon standard for comparison of evaluation results (e.g., the answer key of a school quiz, or the map of a minefield). Therefore we must somehow accommodate some highly subjective judgments about which translation might be better than which other translation.

O estudo conclui que as classes de erros são dependentes de pares linguísticos, por exemplo da geração de tempos verbais para tradução do inglês para o espanhol, e de ordem de palavras para tradução do chinês para o inglês (VILAR et. al, p. 702).

No que se refere à classificação de erros de tradução, Vilar et al. (2006) expandem a tipologia de erros de TA apresentada por Llitjós et al. (2005). Esta expansão se dá nas categorias de “palavras faltantes”, “ordem de palavras”, e “palavras incorretas”²² (essas categorias são detalhadas na seção 3.1). Isso se deve de uma forma geral pelas línguas trabalhadas em cada um dos estudos, já que Llitjós et al. (2005) analisam as traduções automáticas do inglês para o espanhol, e Vilar et al. (2006) analisam os pares linguísticos inglês-espanhol e chinês-inglês em processo de tradução automática.

Tais estudos são necessários para contextualizar a pesquisa na área de ATA em língua portuguesa. Ribeiro (2006) oferece uma revisão das categorias de erros de TA de

²² Tradução livre.

três trabalhos envolvendo a língua portuguesa em processo de tradução automática, a saber as dissertações de Dayrell (1999), Oliveira et. al. (2000), e dois capítulos de livro editados por Diana Santos²³: o de Sarmiento (2007) e Maia e Barreto (2007), os quais são mencionados por Ribeiro (2006) como trabalhos que apresentam as categorias de erros de tradução automática utilizada no TrAva (vide definição na página 77). Embora a finalidade da dissertação (RIBEIRO, 2006) seja analisar um estudo de caso, especificamente no contexto da localização de software (o qual não é levado em consideração no presente estudo), a revisão dessas categorias contribui para estabelecer uma literatura inicial com vistas a pesquisa de categorias de erros de tradução automática envolvendo a língua portuguesa.

Silva (2010) desenvolve uma análise comparativa de resultados de tradução automática baseados em regras (*Systran*) e estatística (*Google Translate*). Utiliza como método a análise manual de erros com base na tipologia de Vilar et al. (2006), Calude (2004), e Ribeiro (2006). O estudo, muito embora utilize alguns contextos de produção para descrever erros de tradução automática, e com base na noção de gêneros textuais de Marcuschi (2002), não se atém a fatores sócio-culturais ou psicológicos, como sugere a área da localização de software. Nesse aspecto, o trabalho de Silva (2010) apresenta

²³ A profa. dra. Diana Santos é consagrada na área de PLN utilizando a língua portuguesa. Possui diversas publicações internacionais sobre as mais diversas áreas da PLN, contando inclusive com trabalhos na área da TA e ATA especificamente. O livro publicado em (SANTOS, 2007) com sua editoria conta com uma seção exclusivamente dedicada a trabalhos na área de TA envolvendo a língua portuguesa.

perfil similar ao do presente estudo.

É notável que no contexto brasileiro, há uma escassez de pesquisa na área interdisciplinar de ATA, especialmente a partir de pontos de vista que vão além das ciências exatas aplicadas, tais como as ciências da computação ou engenharias. As dissertações citadas acima, advindas da área de Letras, quais sejam de Dayrell (1999), Ribeiro (2006), e Silva (2010), fornecem um importante subsídio teórico para preencher essa lacuna, especificamente, no que tange o método chamado de *black box*, ou seja a avaliação externa (do resultado gerado pela tradução) sem o exame do funcionamento do sistema internamente (chamado *glass box*).

Tal lacuna é levantada por Sarmiento et. al. (2007) e corroborada em Ribeiro (2006, p. 39) ao descrever que a “categorização [de erros de TA] deveria ser elaborada por pesquisadores com formação em linguística, tradução humana, e tradução automática, não havendo ainda muitos profissionais com esse perfil diversificado”. O perfil deste pesquisador se insere justamente nessa declaração, qual seja a de um mestre na área de linguística aplicada em inglês, doutorando na ciência da informação e professor de um curso bacharelado de aplicação e tratamento de língua estrangeira em meio digital.

A partir desses trabalhos, situados na interface da tradução e tecnologia, delinea-se o aspecto da avaliação da tradução automática no contexto de investigação

do objeto proposto, qual seja da análise de resultados de tradução automática em documentos multimodais.

A subseção a seguir busca identificar o contexto de investigação da linguística computacional, e mais especificamente da tradução automática dentro de uma amostra de periódicos da área da ciência da informação no Brasil.

2.5 Ciência da Informação e Linguística Computacional

Conforme descrito na introdução deste capítulo, desde o começo da revolução digital nos anos cinquenta, ambas as discussões sobre o estabelecimento da Ciência da Informação enquanto disciplina acadêmica e a pesquisa em tradução automática (enquanto um sistema de aplicação, e não uma disciplina acadêmica) começaram a se tornar uma realidade.

De acordo com a *Série Branca* da METANET (Aliança Tecnológica Europeia Multilíngue) sobre a língua portuguesa na era digital (BRANCO, 2012, p. 72), nota-se um escasso apoio e recursos para a aplicação de tradução automática (TA) envolvendo a língua portuguesa. E, segundo Branco (ibid., p. 25), um dos desafios principais nessa área de aplicação é a adaptação de recursos linguísticos para um domínio ou área específica de uso.

A *Série Branca* (BRANCO, 2012), que neste tópico trata do processamento da língua portuguesa, relata que os estudos sobre a tradução automática no Brasil tem sido mais desenvolvidos em alguns núcleos de pesquisa localizados em São Paulo (especialmente o NILC – Núcleo Interinstitucional de Linguística Computacional), Rio de Janeiro e Rio Grande do Sul, com uma abordagem mais processual, isto é, do processamento da língua natural, sendo assim iluminados pelas áreas mais próximas a aplicações matemáticas e computacionais, tais como ciências da computação e Estatística, e pouco pela representação e formalização da língua para o processamento, ou seja, da linguística computacional, de uma perspectiva mais próxima da linguística e da ciência da informação (CI).

Diversos estudos na interface entre a linguística e a CI já foram realizados e divulgados na literatura da ciência da informação, revelando os pontos de contribuição para a CI como a Terminologia e a Análise Documentária, e problemas mais relacionados à construção de conceitos e representação da informação (ALMEIDA, 2011; MENDONÇA, 2000; MOLLICA; GONÇALEZ, 2012).

Assim, é crucial identificar as relações da linguística computacional e ciência da informação dentro da literatura nacional da última (CI). Dessa forma, torna-se possível verificar o conhecimento produzido nessa interface, oferecendo um panorama no contexto nacional do que já foi pesquisado e publicado.

A grande parte das publicações são artigos de Qualis A1, A2, e B1²⁴ da área das Ciências Sociais Aplicadas (De acordo com o Qualis/CAPES 2013) a partir dos seguintes periódicos:

- *Brazilian Journal of Information Science* - Unesp/Marília
- *Transinformação* - PUC-Campinas
- *Ciência da Informação* – IBICT
- *Perspectivas em Ciência da Informação* - UFMG
- *Incid: Ciência da informação, Informação, Documentação* - USP-Ribeirão Preto
- *Informação e Sociedade* - UFPB
- *Datagramazero* - Rio de Janeiro

Há outros periódicos relevantes para o levantamento de produção nacional no contexto da LC e CI. Contudo os periódicos acima citados são suficientes para se obter

²⁴ De acordo com informação disponibilizada no sítio da CAPES, "Qualis é o conjunto de procedimentos utilizados pela Capes para estratificação da qualidade da produção intelectual dos programas de pós-graduação.[...] A classificação de periódicos e eventos é realizada pelas áreas de avaliação e passa por processo anual de atualização. Esses veículos são enquadrados em estratos indicativos da qualidade - A1, o mais elevado; A2; B1; B2; B3; B4; B5; C - com peso zero. Note-se que o mesmo periódico, ao ser classificado em duas ou mais áreas distintas, pode receber diferentes avaliações. Isto não constitui inconsistência, mas expressa o valor atribuído, em cada área, à pertinência do conteúdo veiculado. Por isso, não se pretende com esta classificação que é específica para o processo de avaliação de cada área, definir qualidade de periódicos de forma absoluta.". Para maiores informações vide <http://www.capes.gov.br/component/content/article?id=2550:capex-aprova-a-nova-classificacao-do-qualis>. Acessado em 10 de novembro de 2016.

uma amostra representativa da produção de conhecimento pela temática da LC e CI.

O critério utilizado para selecionar quais artigos que se configuram na interface da LC e CI foi a análise de título e do resumo, a partir da pesquisa da palavra-chave "linguística computacional" no *website* do periódico.

Este mesmo critério de seleção de palavra-chave de busca foi realizado com base nos resultados encontrados, considerados satisfatórios para uma amostra representativa. De fato, outros artigos relevantes do ponto de vista da temática aqui trabalhada, ficaram de fora por não apresentarem a "linguística computacional no título ou resumo do texto.

Depois de realizar as buscas nos periódicos selecionados, observou-se que de forma geral houve um número relativamente baixo de resultados a partir da busca contendo a palavra-chave "linguística computacional". Seguindo o padrão dos resultados encontrados para a palavra "linguística" no estudo de Mendonça (2000), o maior número de resultados contendo a palavra-chave "linguística computacional" foi da revista *Ciência da Informação* do IBICT, e as revistas *Brazilian Journal of Information Science*, *Transinformação*, e *Datagramazero* não retornaram resultados.

Uma vez realizada a busca pela palavra-chave nos periódicos supracitados, chegou-se aos seguintes resultados:

- *Brazilian Journal of Information Science* - 0 resultado

- *Transinformação* - 0 resultado
- *Ciência da Informação* - 8 resultados
- *Perspectivas em Ciência da Informação* - 1 resultado
- *Incid: Ciência da informação, Informação, Documentação* - 4 resultados
- *Informação e Sociedade* - 1 resultado
- *Datagramazero* - 0 resultado

A partir da listagem acima, percebe-se um maior número de artigos contendo a palavra-chave “linguística computacional” no periódico do IBICT, *Ciência da Informação*, seguido pela revista *Incid*, com metade dos resultados da anterior, finalizando a contagem bastante escassa das revistas *Perspectivas em Ciência da Informação e Informação e Sociedade*, com 1 (hum) resultado apenas cada. Os demais periódicos não apresentaram resultados a partir da busca pela palavra-chave “linguística computacional”. Assim, totalizou-se 14 artigos de maior relevância para a ciência da informação, contendo “linguística computacional” no título ou resumo a partir de busca na base de dados do periódico.

Contudo, esses números não são absolutos. Eles devem ser interpretados com uma orientação para seus contextos de produção. Assim por exemplo, observamos um maior número na revista *Ciência da Informação*, pois além, da temática linguística presente em grande parte dos trabalhos (MENDONÇA, 2000), a base de dados é vasta

(desde 1972 até 2014), e a publicação é quadrimestral (ou seja, três vezes ao ano).

Outro aspecto importante é da representação da palavra-chave “linguística computacional” com conceitos sinônimos. Diversos textos, além desse quantitativo, discutem ou citam a linguística computacional com outras palavras, tais como o artigo de Diana Santos (SANTOS, 2002), uma sumidade na área da linguística computacional da língua portuguesa, para a *Datagramazero*, “Um centro de recursos para o processamento computacional do português”. Embora não haja a expressão “linguística computacional” no título, trata-se da mesma temática.

A amostra demonstra que a menção à área é de fato ainda pouco expressiva na literatura da ciência da informação no Brasil. Porém, apresenta uma certa trajetória e tendência ao longo do tempo. Por exemplo, na revista *Ciência da Informação*, editada pelo IBICT, em produção desde 1972 até o último número deste ano, foi possível encontrar artigos desde 1983 até 2009 mencionando a área da LC.

As tabelas abaixo demonstram os periódicos, datas, títulos, autores e instituições para uma melhor visualização dos dados.

Tabela 1 - Linguística computacional na revista *Ciência da Informação*

	Periódico	Data	Título	Autor	Instituição
1.	Ciência da Informação	2009	A pragmática no contexto da	Rodrigues, J.; Caricatti, A.	UnB; UnB

			identificação de autoria de textos		
2.	Ciência da Informação	2009	O poder cognitivo das redes neurais artificiais modelo na recuperação da informação	Capuano, E.	UnB
3.	Ciência da Informação	2007	A diversidade lingüística da Internet como reação contra-hegemônica das tendências de centralização do império	Guessser, A.	Universidade de Coimbra
4.	Ciência da Informação	2006	Aplicación de transductores de estado-finito a los procesos de unificación de términos	Galvez, C.	Universidade de Granada

5.	Ciência da Informação	2004	Novas fronteiras tecnológicas das ações de informação: questões e abordagens	González de Gomez, M.	Ibict-MCT
6.	Ciência da Informação	1995	O léxico na economia da língua	Correia, M.	Universidade de Lisboa
7.	Ciência da Informação	1991	Indexação automática de textos: uma abordagem otimizada e simples	Robredo, J.	UnB
8.	Ciência da Informação	1983	Perspectivas na contribuição da linguística e de áreas afins à ciência da informação	Baranow, U.	UnB

Fonte: Próprio autor.

Tabela 2 - Linguística computacional na revista *Perspectivas em Ciência da Informação*

Periódico	Data	Título	Autor	Instituição
Perspectivas em Ciência da Informação	2005	SiRILiCO uma proposta para um sistema de recuperação de Informação baseado em teorias da linguística computacional e ontologia	Duque, C. D.	UFMG

Fonte: Próprio autor.

Tabela 3 - Linguística computacional na revista *Incid: Ciência da Informação, Informação, Documentação*

	Periódico	Data	Título	Autor	Instituição
1.	Incid: Ciência da informação, Informação, Documentação	2013	Luciano Floridi e os problemas filosóficos da informação: da	González de Gomez, M.	UFF

			representação à modelização		
2.	Incid: Ciência da informação, Informação, Documentação	2010	Parâmetros teóricos para elaboração de instrumentos pragmáticos de representação e organização da informação na Web: considerações preliminares sobre uma possível proposta metodológica	Gracioso, L.	UFSCar
3.	Incid: Ciência da informação, Informação, Documentação	2010	Deleuze e Guattari e a Psicologia Cognitiva, IA e IHC: investigando	Day, R.	Indiana University

			possíveis conexões e diferenças		
4.	Incid: Ciência da informação, Informação, Documentação	2010	A análise do discurso e o campo informacional: usos atuais e alcance epistemológico: uma atualização	Freitas, L.	UFF

Fonte: Próprio autor.

Tabela 4 - Linguística computacional na revista *Informação e Sociedade*

Periódico	Data	Título	Autor	Instituição
Informação & Sociedade	2008	INDEXAÇÃO AUTOMÁTICA E SEMÂNTICA: estudo da análise do conteúdo de teses e	Borges, G.; Maculan, B.; Lima, G.	UFMG; UFMG; UFMG.

		dissertações		
--	--	--------------	--	--

Fonte: Próprio autor.

Em relação a trajetória e tendência, podemos perceber que a menção à LC na CI ganhou mais destaque na última década: a maior parte dos artigos das revistas analisadas foram publicados a partir de 2004. A pequena parcela de artigos datados do século passado já aponta para essa interface desde 1983, passando por um longo intervalo até emergir novamente o tema na década de noventa.

Sobre os títulos dos artigos dos periódicos, pode-se observar a configuração de determinadas categorias ligadas à linguística e mais especificamente à linguística computacional. E em um recorte mais diacrônico, nota-se uma tendência de aprofundamento da interação entre áreas mais específicas da linguística e da CI a partir do artigo de Baranow (1983) para a temática mais próxima à linguística computacional ou o processamento de língua natural e a CI²⁵.

Em relação aos autores, verificamos que há uma variedade de autores de diversas áreas e instituições que contribuem para a pesquisa em ciência da informação. Portanto, fica clara a natureza interdisciplinar da área.

A partir do esboço quantitativo-panorâmico de uma amostra representativa do

²⁵ Tais como a indexação automática, recuperação da informação, pragmática, e análise de conteúdo.

conhecimento produzido no Brasil acerca das relações entre a linguística computacional e a ciência da informação, cruza-se esses dados com os conteúdos dos artigos, buscando obter uma configuração das principais temáticas que emergem no contexto brasileiro.

A partir dos artigos analisados, observa-se que há diferentes tipos de abordagens acerca das relações entre a linguística computacional e a CI. Alguns inclusive se distanciando mais da noção "intrínseca" de processamento ou formalização da língua natural no computador.

O item 3 da Tabela 1 apresenta uma visão mais sociológica (também por conta de sua formação na sociologia) da diversidade linguística presente na Internet, e não apresenta o termo linguística computacional no texto. O item 1 da tabela 1, embora traga a palavra-chave "linguística computacional" no resumo, não há qualquer relação direta no texto nesse sentido, mas sim uma ênfase de duas categorias linguísticas para a identificação de estilo de autoria.

Já o item 8 dá mais ênfase ao aspecto linguístico em sua interface com a ciência da informação. Contudo, neste artigo, Baranow (1983) apresenta a linguística computacional como uma subárea especializada para a formação do cientista da informação. O autor (ibid.) propõe uma subseção inteiramente dedicada à contribuição da CI para a formação do cientista da informação. Baranow (1983, p. 31-31) identifica três áreas de atuação nesse sentido: a) processamento automático de texto; b)

recuperação automática da informação; c) sistemas automáticos de pergunta-resposta (inteligência artificial).

Muito embora o desenvolvimento tecnológico tenha avançado substancialmente nos últimos tempos, observa-se também o desenvolvimento dessa áreas nos artigos mais recentes. Assim, na tabela 1, item 2, percebemos a combinação de duas áreas observadas por Baranow (1983), quais sejam da recuperação da informação e inteligência artificial. Já no item 4, Galvez (2006) aprofunda o que seria a área a) de Baranow (1983), processamento automático de texto, em seu trabalho de avaliação de lematizadores (programas de recuperação de informação que reduzem variações de termos semanticamente equivalentes) na unificação de termos.

O item 6 da tabela 1 trabalha em um sentido mais aplicado da linguística computacional como o item 4, porém prioriza a gestão e geração de significado lexical, por meio da formalização e representação linguística (em oposição a modelos matemáticos ou estatísticos). O item 5 discute a linguística computacional de forma parecida a Baranow (1983), em uma perspectiva mais teórica, porém de forma mais ampla (agregando uma perspectiva mais filosófica e social nessa interface) ampliando a relação de linguagem, informação e comunicação em duas grandes linhas, nas quais, em uma das linhas a linguística computacional e processamento de língua natural são alocadas, enquanto concepção da "linguagem como dimensão dos dispositivos de

tratamento da informação” (GONZALEZ DE GOMEZ, 2004, p. 57). Por fim, o item 7 trabalha com a abordagem da LC para resolver problema de palavras vazias na indexação automática de um sistema de recuperação de informação.

Podemos perceber essa mesma configuração nos outros artigos. Na tabela 3, por exemplo, no item 1 há uma repetição de autora (GONZALEZ DE GOMEZ, 2004, 2013) reforçando a linha filosófica no tratamento automático da informação. O item 2 da tabela 3 cruza diversas características de pesquisa na interface, na utilização de vocabulário controlado e folksonomias (indexação colaborativa na *web*) para recuperação da informação. Mas com o viés da utilização da interface para propor uma metodologia, de forma similar ao item da tabela 2 que utiliza a área da linguística computacional juntamente com a ontologia como abordagens teórico-metodológicas para criar um sistema de recuperação de informação, chamado de SIRILICO (DUQUE, 2005) (expansão das três áreas citadas por Baranow em 1983).

O item 3 da tabela 3 relaciona a perspectiva cognitiva da inteligência artificial e interação homem-máquina com o item 2 da tabela 1. E o item 4 da tabela 3 fica mais longe do núcleo computacional da interface analisada, permanecendo mais próximo do item 1 da tabela 1 por tratar com profundidade de uma área da linguística, qual seja a da análise do discurso, e seus pontos de intersecção com a linguagem, informação e comunicação proposta por Gonzalez de Gomez (2004). Por fim, a tabela 4 aponta para a

indexação automática em diálogo com problemáticas semânticas, conforme, por exemplo, os itens 7 e 8 da tabela 1.

A presente seção identificou relações da linguística computacional e ciência da informação dentro da literatura nacional da última (CI). O levantamento bibliográfico de sete periódicos de relevância para a área apontam para quatorze artigos a partir da pesquisa da palavra-chave “linguística computacional”. Nota-se que desses quatorze artigos, uma pequena parte não trata de fato da linguística computacional.

Nota-se temas em comum, mesmo em diferentes revistas, ressaltando áreas de intersecção como a recuperação da informação, o processamento automático de texto, e inteligência artificial. Observa-se linhas mais teóricas, utilizando uma perspectiva mais filosófica e sociológica para se debruçar sobre o tratamento automático da língua em sua interação com a informação, e outras mais aplicadas, por exemplo propondo a linguística computacional como base teórico-metodológica para o desenvolvimento de um sistema de recuperação da informação (DUQUE, 2005). Por fim, identifica-se um grande interesse pelas áreas de aplicação de busca, e raras menções sobre a aplicação de tradução automática.

A próxima seção aborda a perspectiva da multimodalidade, que permeia o contexto de investigação proposto neste trabalho.

2.6 Multimodalidade

A multimodalidade é uma abordagem de caráter interdisciplinar baseada em teoria comunicativa e semiótica social que caminha em direção a uma teoria (KRESS; VAN LEEUWEN, 2006). Linguagem além da linguagem, segundo alguns estudiosos da área, a multimodalidade preocupa-se com o entendimento e representação isoladamente e interconectadamente das formas comunicacionais que as pessoas utilizam, tais como gesto, postura, imagem e outras formas (JEWITT, 2009).

Jewitt (2009, p.14-16, tradução nossa) define quatro pressupostos que subsidiam uma aceção em geral compartilhada entre os estudiosos da multimodalidade:

- A língua é parte de um conjunto multimodal. Isso quer dizer que a multimodalidade entende que a representação e comunicação sempre influenciam numa multiplicidade de modos (do inglês *mode* - por exemplo, gestos, posturas, imagens, e sons), todos com potencial para contribuir igualmente para um significado;
- Entende-se que cada modo em um conjunto multimodal constrói diferentes trabalhos comunicativos. A multimodalidade supõe que, como a língua, todos os modos tomaram forma por meio de seus usos sociais, culturais e históricos para construírem funções sociais. Portanto, imagens e outros modos não-linguísticos tem um papel específico, em um contexto específico, e em um momento específico;
- Indivíduos orquestram significado por meio da seleção e configuração de modos. Dessa forma, a interação entre os modos é importante para se fazer significado. Os significados em qualquer modo estão sempre interligados com os significados feitos com aqueles de todos os outros modos co-presentes e co-operantes no evento comunicativo;

- Os significados dos sinais modelados a partir de recursos semióticos multimodais são, como a fala, sociais. Isto é, eles são moldados pelas normas e regras que operam no momento da construção do sinal, influenciados pelas motivações e interesses do sinalizador em um contexto social específico. Isto é, produtores de sinais selecionam, adaptam e remodelam significados por meio do processo de leitura/interpretação do sinal.

Os pressupostos descritos por Jewitt (2009), representam uma convergência entre alguns autores, muito embora haja outras relações em relação ao uso de modo, expandindo ou modificando seu significado dependendo de cada contexto de uso, conforme Pires e Duque (PIRES; DUQUE, 2015) destacam:

[...]é importante o entendimento de cada um desses termos, tais como modo, recursos semióticos, modalidade são articulados no determinado contexto de ocorrência para observarmos sua articulação e manifestação em um contexto socialmente e culturalmente construído. Por exemplo, Kress (KRESS, 2009, p. 54) define multimodalidade como um “mapeamento de um domínio de investigação”, dado as singularidades quando aplicada em áreas diferentes, com problemas distintos, como Medicina, Antropologia, e Educação, por exemplo. Nesse sentido, percebe-se que a multimodalidade segue o desenvolvimento científico com vistas para a realidade complexa de um mundo cada vez mais globalizado, onde diferentes modos semióticos são utilizados para disseminar mensagens e articular diferentes mídias na produção de significado com potencial de entendimento que extrapola as barreiras geográficas. Dessa forma o entendimento de como as tecnologias linguísticas são utilizadas para atingir esses objetivos é relevante para o estudo do fenômeno informacional entre diferentes línguas e culturas.

Kaltenbacher (2004) faz uma revisão dos principais trabalhos que contribuíram para a elaboração da multimodalidade enquanto área de pesquisa. O autor investiga as primeiras análises de diferentes modos semióticos realizadas por alemães classicistas e

apresenta características de uma análise similar à multimodal, porém sem utilizar os mesmo termos.

O autor (KALTENBACHER, 2004) também descreve as principais áreas que subsidiaram a abordagem como é conhecida nos últimos vinte anos. Sendo a teoria linguística advinda dos conceitos da semiótica social de Halliday (1975, 1978), precursor da linguística sistêmico-funcional²⁶, que influenciou substancialmente o estabelecimento da abordagem da multimodalidade conforme conhecida recentemente (reconhecida pelo trabalho pioneiro de Kress e van Leeuwen em 1996 e O'Toole (1994) em 1994).

Kress e van Leeuwen (2001, p.1) procuram fornecer uma “terminologia comum para todos os modos semióticos”, diferente do livro seminal da área, *Reading Images* (KRESS; VAN LEEUWEN, 1996), quando se propuseram a compilar uma gramática para o visual, observando apenas um tipo de modo (modo visual tal como imagens). Kress e Van Leeuwen (2001, p.1) aplicam uma “terminologia comum para todos os modos semióticos em um dado domínio sociocultural, [assim] os mesmos significados podem ser frequentemente expressados em diferentes modos semióticos”.

Este novo conceito deixa de lado a ideia de um papel especial fixado para cada

²⁶ vide Halliday, (1975, 1978), Iedema (,2003), Halliday e Hasan (1976), e De Beaugrande e Dressler (1981) respectivamente.

modo, como por exemplo a ideia de que música deve ser somente interpretada em termos de sons e emoção. Dessa forma, tal conceito multimodal considera “princípios semióticos comuns operando dentro e através de diferentes modos” para “ser possível para a música codificar ações ou para imagens codificarem emoções” (KRESS; VAN LEEUWEN, 2001, p.2).

Para os autores Kress e Van Leeuwen (2001, p.2), na era da digitalização os diferentes modos têm se tornado tecnicamente os mesmos em algum modo de representação, e eles podem ser operados por uma pessoa com multi-habilidades, por meio do uso de uma interface, um modo de manipulação física, possivelmente indagando “devo expressar isso com som de música?, devo dizer isso visualmente ou verbalmente?”.

Observa-se que nessa perspectiva, o elemento “unificante e unificado” da tecnologia e semiótica desenha uma linha em direção ao discurso multimodal, pois para o discurso acontecer ele deve conter elementos coesivos que “se interligam” para formar uma ideia coerente e unificada, via interação de diferentes modos semióticos em um dado domínio social.

Linguistas tradicionais trabalham com a ideia de “dupla articulação”, isto é quando textos são articulados como forma e como um significado. No entanto, para Kress e van Leeuwen (2001, p.2) os significados articulados do texto multimodal

multiplicam-se. E estes significados articulam-se em quatro domínios de prática chamados "estratos"²⁷. Esses quatro estratos são o discurso, design, produção e distribuição.

Para Kress e van Leeuwen (2001, p.3), a definição de discurso é de um "conhecimento socialmente construído da, ou de algum aspecto da realidade". Eles explicam que algo "construído socialmente" ocorre em vários "contextos específicos e de uma maneira "apropriada ao interesse de atores sociais nesses contextos, que podem ser amplos como a cultura ocidental, ou exíguo como uma conversa entre irmãos (exemplos meus).

Outra vertente da multimodalidade que adota o discurso multimodal, porém dentro de uma perspectiva empírica é a do grupo de trabalho de John Bateman. Seu trabalho mais conhecido na área, "Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents" (Bateman, 2008) elabora uma metodologia consistente para a análise empírica de documentos multimodais, ou seja, onde há "interação" e "combinação" de múltiplos modos dentro de um mesmo artefato.

Para Bateman (2008, p.1) "modo" é todo aspecto visual diverso de se apresentar informação. Assim, o autor explica que:

A combinação desses modos dentro de um único artefato - no caso do impresso, que seja encadernado, grampeado, ou dobrado, ou mídia online, por

²⁷ Do inglês, *strata* - termo adaptado da linguística sistêmico-funcional de Halliday.

meio de “links” com uma variedade de “hiperlinks” - traz a tona nosso principal objeto de estudo: os documentos multimodais.

Além da definição de modo, o outro elemento que fundamenta sua proposta metodológica é a noção de gênero. O autor (Bateman, 2008, p.9-11) atribui descrições que fundamentam as noções de gênero utilizado na análise de documentos multimodais, a saber i) a noção informal de gênero tais como "gênero de websites", "gênero de jornais" e o sentido onde esses gêneros são realizados; ii) gênero permite teorizar sobre uma gama de possibilidades abertas aos documentos²⁸; e iii) considerar a materialidade dos artefatos multimodais (documentos incluídos) como componente crucial na concepção de gênero multimodal.

Bateman (2008) elabora um recurso sistemático e analítico (originalmente chamado de *GeM project*) de tais documentos baseando-se em "camadas" chamadas de gênero, navegação, *layout*, e estrutura retórica. A figura a seguir demonstra essas camadas utilizadas para identificar os níveis de interação e combinação de diferentes aspectos para posterior manipulação (por meio de programação):

Tabela 5 - Camadas principais da estrutura conceitual de Gênero e Multimodalidade

²⁸ Neste trecho o pesquisador exemplifica ao caracterizar gênero não como uma coleção “solta” separada por tipos de textos, mas como “pontos”, ou melhor “regiões”, já que gêneros, quando observados em detalhe, podem mudar, hibridizar com, ou colonizar, um com outro (Bateman, 2008, p.10).

<i>Content structure</i>	the content-related structure of the information to be communicated—including propositional content
<i>Genre structure</i>	the individual stages or phases defined for a given genre: i.e., how the delivery of the content proceeds through particular stages of activity
<i>Rhetorical structure</i>	the rhetorical relationships between content elements: i.e., how the content is 'argued', divided into main material and supporting material, and structured rhetorically
<i>Linguistic structure</i>	the linguistic details of any verbal elements that are used to realize the layout elements of the page/document
<i>Layout structure</i>	the nature, appearance and position of communicative elements on the page, and their hierarchical interrelationships
<i>Navigation structure</i>	the ways in which the intended mode(s) of consumption of the document is/are supported: this includes all elements on a page that serve to direct or assist the reader's consumption of the document

Table 1.2 The primary layers of the Genre and Multimodality framework

Fonte: Bateman (2008, p. 19)

A tabela acima foi retirada de Bateman (2008, p.19). Ela demonstra as camadas e correspondentes definições que orientam o projeto de gênero e multimodalidade para a análise de documentos multimodais.

A seguinte subseção revisa algumas pesquisas recentes relacionando a multimodalidade e o uso de tecnologias dentro do contexto da CI no Brasil.

2.7 Multimodalidade e Ciência da Informação

A interação entre a Multimodalidade e a Ciência da Informação no Brasil é ainda

muito recente. Porém a interface já conta com algumas produções finalizadas e outras em andamento, especificamente dentro do grupo de pesquisa chamado R.E.G.I.I.M.E.N.T.O.²⁹ (Arquitetura da Informação, Linguística Computacional e Multimodalidade, Mídia e Interatividade), liderado pelo professor dr. Claudio Gottschalg-Duque. O Diretório de Grupos de Pesquisa do CNPq recentemente certificou o grupo e é o único que trabalha com esta área interdisciplinar, no contexto nacional.

Dentre as produções mais recentes, destacam-se as seguintes:

- A tese de Sampaio (2016) busca definir uma abordagem metodológica que possibilite outros instrumentos, além dos já existentes, para a definição das categorias espaciais na construção de mapas com o objetivo de ampliar seu uso a partir de variados tipos de modos. Essa investigação tem como modelo conceitual a ciência da informação geográfica, a teoria da relevância, partindo do princípio do menor esforço no processo comunicacional e a teoria espacial pós-moderna como bases teóricas e a multimodalidade como abordagem metodológica para a construção de uma arquitetura

²⁹ O objetivo deste grupo é pesquisar o estado da Arte da Arquitetura da Informação, Linguística Computacional, Ontologias e Multimodalidade, identificando as vantagens e desvantagens do ponto de vista de aplicação prática de teorias destas áreas e propor a utilização conjunta de conhecimento, tanto no desenvolvimento / adaptação de material acadêmico / de ensino especificamente, mas não limitado a, tais como produtos interativos virtuais e educacional / de ensino, tais como livros digitais, a educação virtual. O Grupo já produziu um livro que se encontra atualmente no II (em português) e III (em inglês) volumes, bem como vários artigos em periódicos e eventos de diferentes áreas do conhecimento humano (basicamente, Ciência da Informação, Ciência da Computação e Pedagogia)

multimodal da geoinformação. O objetivo é propor uma metodologia para definição de simbologias de representação de dados geoespaciais com base nas classes de objetos da ET-EDGV para os dados cadastrais, a partir de uma arquitetura da geoinformação multimodal associada ao método de categorização cognitiva da geoinformação a ser visualizada nos mapas a fim de ampliar a capacidade de comunicação cartográfica.

- A dissertação de mestrado de Santos (2013) propõe verificar os possíveis benefícios provenientes do uso da Interação Humano-Computador Multimodal (IHCM) em uma interface computacional baseada em voz artificial - voz gerada por computador associada a um mecanismo de recuperação da informação em que se busca a melhoria do diálogo entre o homem e a máquina nas operações de troca de informação.

- A tese de doutorado de Sousa (2015) busca analisar a percepção do aluno para gerar informação relevante ao professor em forma de feedback. Para tanto, propõe uma modelagem inovadora com aspectos da teoria da atividade, multimodalidade e uso de ontologia. A modelagem busca usar o estudo do comportamento humano e como a consciência é formada por meio da conceituação de atividade proposta por Vygostky (1978), que permite discutir sobre as ações, motivos e objetivos das atividades educacionais. Para análise das atividades multimodais executadas por um professor dentro do ambiente educacional, recorreu à abordagem multimodal proposta por Hodge e Kress (1988), que proporciona a discussão no seu viés sócio-construtivista. O

foco do estudo na sala de aula do ensino técnico traz a necessidade da compreensão da semiótica da comunicação multimodal, pelo uso regular de ação e imagens na sala de aula. Como resultado, criou-se um modelo de organização da informação nas salas de aulas que utiliza uma ontologia com 45 atividades do professor que podem ser auxiliadas com a percepção dos alunos no ensino técnico do Distrito Federal. Além disso, apresentou-se um modelo automático que gera mais de 68 bilhões de feedbacks diferentes.

- A tese de Steinmetz (2015) apresenta uma proposta de contribuição da Arquitetura da Informação na construção de Ambientes Informacionais Colaborativos de Ensino-Aprendizagem com a intenção de estabelecer um fluxo de atividades para elaboração, utilização e refinamento desses ambientes, a partir dos conceitos e fundamentos da arquitetura da informação, da multimodalidade e da Teoria da Relevância. Para tanto, realizou uma revisão bibliográfica a fim de identificar o estado da arte dos temas envolvidos na pesquisa. Em seguida, identificou estudos sobre a relação entre a ciência da informação, a arquitetura da informação e a multimodalidade aplicadas em Ambientes Informacionais Colaborativos de Ensino-Aprendizagem, com o objetivo de aumentar a relevância desses ambientes para os estudantes. Elaborou uma proposta inicial do fluxo do modelo de arquitetura da informação multimodal aplicado em ambientes Colaborativos de Ensino/Aprendizagem.

2.8 Considerações Finais

A partir das subseções acima nota-se que há um contexto interdisciplinar no qual o problema da presente pesquisa se insere, envolvendo a CI, a LC e a Multimodalidade e suas interrelações. A discussão levantada neste capítulo é proporcionalmente relevante por conta do caráter exploratório do estudo. E assim, fornece o contexto para que se possa entender os aspectos teóricos que informam o percurso investigativo.

3. Marco Teórico

O capítulo anterior tratou de discutir a literatura que contextualiza o presente estudo. É dentro desse contexto que este capítulo descreve o aparato teórico que subsidia e informa a análise do problema que emerge a partir do estudo piloto. Tal descrição se realiza por meio de duas subseções: a primeira sobre os estudos de categorias de erros de tradução automática, e a segunda sobre os estudos de relações intersemióticas.

3.1 Categorias de erro de tradução automática

No contexto de estudos sobre ATA constata-se duas principais metodologias de investigação: uma que permite examinar o funcionamento por dentro da máquina (chamada de *glass box*), e a outra onde é possível acessar os textos de entrada (original) e de saída (tradução), sem levar em consideração a engenharia do processo, chamada de *black box*.

White (2003) reúne essas informações para demonstrar formas de se realizar a avaliação de tradução automática. Como a perspectiva levada em consideração neste estudo se debruça sobre o fenômeno intersemiótico, qual seja aquele entre o resultado de tradução automática que se relaciona semanticamente com uma imagem (ou parte

dela), considera-se apenas o método *black box*³⁰.

White (2003, p. 225) utiliza os seguintes exemplos para ilustrar a vantagem de se “medir a cobertura do sistema, e até obter uma hipótese sobre como o sistema tenta lidar com os fenômenos”:

Figura 1 - Exemplo de perspectiva de avaliação *black box* de um sistema de TA

- (4) a. *There is a gun in my bedroom.*
 b. *Hay un revólver en mi alcoba.*

This looks all right; *there is* is appropriately translated by the Spanish existential copula *haber* (inflected as *hay*). However, in (5),

- (5) a. *Is there a gun in my bedroom?*
 b. **¿Es allí un revólver en mi alcoba?*

it fails, hinting that this system only gets *haber* when the input is exactly *there is* or *there are*. This suspicion is confirmed by (6), where we do not want *there are* to be translated this way, because the two words belong to different constructions.

- (6) a. *Some of the people over there are Spanish.*
 b. **Alguna de la gente sobre hay Español.*

Fonte: White (2003, p. 225)

Conforme a leitura do autor (White, 2003, p. 225) no exemplo 4 acima, a frase 4(a) de entrada em inglês *there is* é traduzida de forma apropriada pelo verbo de ligação *haber*, flexionado como *hay* na tradução gerada automaticamente para o espanhol 4(b). Porém, no exemplo número 5 há uma falha, o que sugere que o sistema reconhece *haber* somente quando o *input* (texto de entrada) está exatamente na ordem *there is* ou *there*

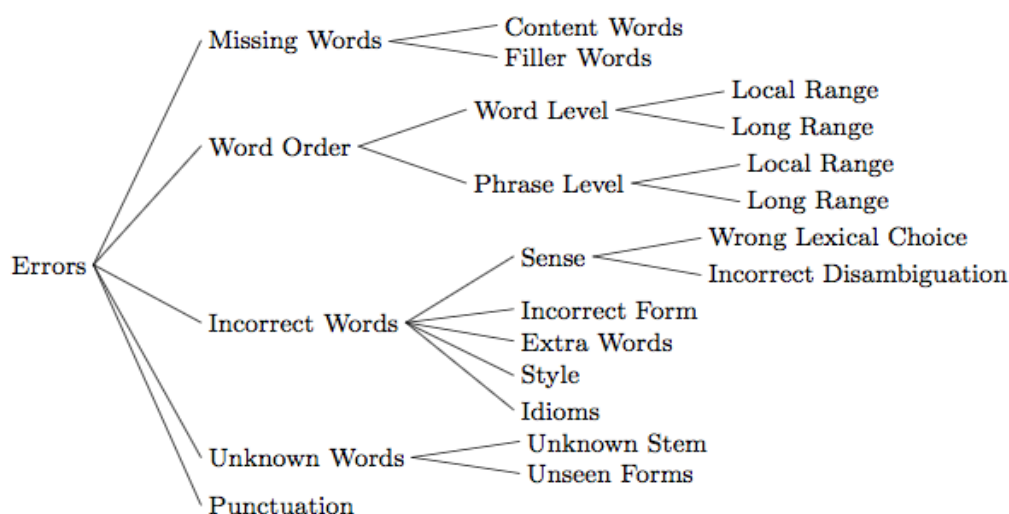
³⁰ Por isso não haveria sentido buscar a parceria com o *Google research program* (<http://research.google.com/>) para investigar a engenharia (*glass box*) do Google Tradutor.

are. A suspeita é confirmada no exemplo número 6, pois a construção de *there are* nesta frase é diferente (ibid.).

Este tipo de exemplo de perspectiva *black box* também recebe um aprofundamento no sentido de utilizar tipologias de erros mais frequentes encontradas em contrastes de entrada e saída de sistemas de tradução automática. E alguns autores, tais como Vilar et. al (2006), utilizam diferentes terminologias para abordar fenômenos dos chamados “erros” de tradução automática (erros no sentido de máquina e sistemas).

No contexto de estudo de tipologia de erros de tradução automática, o trabalho de Vilar et. al (2006) apresenta uma quadro para a classificação de erros de um sistema de máquina de tradução. Essa classificação é uma extensão do trabalho de Llitjós et. al. (2005) e apresenta cinco categorias principais conforme a figura abaixo:

Figura 2 - Classificação de erros de TA



Fonte: (VILAR et al., 2006, p. 699).

Como pode-se notar na figura acima, os autores (VILAR et al., 2006, p. 698) propõem uma classificação de erros de TA dividida em cinco grandes categorias, a saber *missing words* (palavras faltantes), *word order* (ordem de palavra), *incorrect words* (palavras incorretas), *unknown words* (palavras desconhecidas), e *punctuation* (pontuação).

De acordo com Vilar et. al. (2006, p. 698) a primeira categoria representa casos em que falta uma palavra em frases resultantes de uma TA. E as duas subcategorias, *content words* e *filler words*, respectivamente "palavras de conteúdo" que são necessárias para expressar o sentido da frase e "palavras de preenchimento" que seriam necessárias para formar a frase gramaticalmente, porém o sentido é preservado. A segunda categoria diz respeito às reordenações de palavras e locuções. A diferença entre os dois níveis é o movimento de palavras individualmente ou em blocos de palavras na geração das frases. Em relação ao alcance local ou alcance longo (*local range* e *long range*) a distinção não se dá em termos absolutos³¹, mas reside na necessidade de reorganizar as palavras em um contexto local (dentro de um mesmo bloco sintático) ou de mover as palavras para outro bloco.

A seguir, Vilar et. al. (2006, p. 698) descrevem a maior categoria de seu esquema,

³¹ É importante destacar que os autores chamam atenção para o fato de que "a classificação de erros de um sistema de máquina de tradução não é definitivamente absoluta" (VILAR et al., 2006, p. 698, tradução nossa).

a de erros de "palavras incorretas", que podem ser identificadas quando o sistema não encontra a tradução correta de uma palavra. Na primeira subcategoria, a palavra incorreta interfere no significado da frase. E aqui há outras duas subcategorias em relação a desambiguação: em uma, o sistema escolhe uma tradução incorreta (*wrong lexical choice*) e em outra o sistema não consegue desambiguar o significado correto de uma palavra da língua-fonte (*incorrect disambiguation*). A outra subcategoria de palavras incorretas são os erros de "formas incorretas", ou seja, quando o sistema não produz a forma correta de uma palavra, embora a tradução da forma básica seja correta. A subcategoria seguinte se caracteriza por apresentar palavras adicionais na frase gerada. As duas últimas categorias, de menor relevância, se caracterizam pela escolha "ruim" de palavras na tradução, mas que o significado é preservado, embora Vilar et. al. (2006) não considerem completamente correto (estilo), como por exemplo a repetição de palavra em um contexto próximo; e a outra subcategoria se caracteriza pelas expressões idiomáticas que o sistema não reconhece e gera a tradução de um texto "normal"³².

A quarta categoria apresentada por Vilar et. al. (2006, p. 698) é a de palavras desconhecidas (*unknown words*), que podem ser distinguidas entre palavras realmente

³² É importante salientar que embora os autores (VILAR et. al., 2006) destaquem que a classificação de erros não sejam termos absolutos e inambíguos, há vários termos absolutos e carregados semanticamente empregados para se qualificar as traduções, como por exemplo "correta", "ruim", e normal.

desconhecidas (ou “radical”, no inglês *stem*) e formas desconhecidas de radicais conhecidos. A quinta categoria, “pontuação” não é levada em consideração no trabalho apresentado pelos autores, tampouco no presente estudo.

Tais descrições contribuíram para a construção do aporte teórico que foi levado a cabo para se delinear o caminho investigativo deste estudo, mais especificamente para a análise dos dados.

O trabalho desses autores também é relacionado mais especificamente na utilização da língua portuguesa como um dos pares linguísticos em um processo de tradução automática. Ainda dentro da área de ATA, um dos trabalhos mais notórios nesse contexto é o do TrAva (Traduz e Avalia³³). Segundo Santos et. al. (2004, tradução nossa):

O TrAva é um sistema cujo objetivo é abordar alguns dos critérios empregados intuitivamente no julgamento de traduções, por meio da produção de um arcabouço relativamente fácil para se agrupar cooperativamente centenas de exemplos classificados de acordo com os problemas de (máquina de) tradução.

Algumas categorias de erros de tradução podem ser identificadas e reunidas no trabalho de Ribeiro (2006). O trabalho de Ribeiro (2006) reúne algumas categorias de classificação de problemas morfológicos, léxicos, e sintáticos com base no trabalho de

³³ <http://www.linguateca.pt/TrAva/>

Sarmento ³⁴ (2007). As categorias elencadas por Sarmento (2007) podem ser visualizadas nas tabelas abaixo:

Tabela 6 - Classificação de problemas morfológicos e léxicos do TrAva

Morfologia e Léxico	
<p>1. Substantivos</p> <p>a. Escolha Lexical</p> <p>b. Plural dos Substantivos</p> <p>c. Plural dos Substantivos Compostos</p> <p>d. Grau dos Substantivos</p> <p>2. Adjectivos</p> <p>a. Escolha Lexical</p> <p>b. Plural dos Adjectivos</p> <p>c. Adjectivos Uniformes</p> <p>d. Grau dos Adjectivos - Comparativo</p> <p>e. Grau dos Adjectivos - Superlativo</p> <p>3. Advérbios</p> <p>a. Escolha Lexical</p> <p>b. Grau dos Advérbios - Comparativo</p> <p>c. Grau dos Advérbios - Superlativo</p> <p>d. Locuções adverbiais</p> <p>4. Determinantes</p> <p>a. Escolha Lexical</p> <p>b. Artigos</p> <p>c. Outros Determinantes</p> <p>5. Pronomes</p> <p>a. Escolha Lexical</p> <p>b. P. Pessoais</p> <p>c. P. Pessoais (Função Sujeito)</p> <p>d. P. Pessoais (Função Predicativa)</p> <p>e. P. Pessoais (Função OD)</p> <p>f. P. Pessoais (Função OI - forma átona)</p> <p>g. P. Pessoais (Função OI - forma tónica)</p> <p>h. P. Pessoais (Função Objecto de circunstância)</p>	<p>(continuação)</p> <p>i. P. Pessoais (Forma combinada OI + OD)</p> <p>j. P. Possessivos</p> <p>k. P. Demonstrativos</p> <p>l. P. Relativos</p> <p>m. P. Interrogativos</p> <p>n. P. Indefinidos</p> <p>6. Numerais</p> <p>a. Escolha Lexical</p> <p>b. Cardinais</p> <p>c. Ordinais</p> <p>7. Preposições</p> <p>a. Escolha Lexical</p> <p>b. Contractação com Artigos</p> <p>c. Contractação com Pronomes</p> <p>d. Locuções Prepositivas</p> <p>8. Verbos</p> <p>a. Escolha Lexical</p> <p>b. Tempos Simples</p> <p>c. Tempos Compostos</p> <p>d. Modo - poder, dever, ter de/que</p> <p>e. Voz Passiva</p> <p>f. Verbos Reflexivos</p> <p>g. Locuções Verbais</p> <p>9. Conjunções</p> <p>a. Escolha Lexical</p> <p>b. Conjunções Coordenativas</p> <p>c. Conjunções Subordinadas</p> <p>d. Locuções Conjuncionais</p>

Fonte: Sarmento, 2007 apud Ribeiro 2006, p. 60.

³⁴ Na data da publicação da dissertação de Ribeiro (2006), Sarmento (2007) ainda estava no prelo. Não foi possível no entanto obter acesso ao texto original do autor. Por esse motivo, utiliza-se a fonte de Ribeiro (2006).

Tabela 7 - Classificação de problemas sintáticos do TrAva

Sintaxe	
<p>1. Concordância</p> <p>a. Género no interior do SN</p> <p>b. Género entre o N e o Part. Pass.</p> <p>c. Número no interior do SN</p> <p>d. Número entre o SN e o Verbo</p> <p>e. Número entre o N e o Part. Pass.</p> <p>2. Ordem das Palavras</p> <p>a. Interior do SN</p> <p>b. Interior do SV</p> <p>c. Interior do SP</p> <p>d. Interior da Oração</p> <p>e. Interior da Frase</p> <p>3. Elisão</p> <p>a. Artigo em início de frase</p> <p>b. Artigo com nomes próprios</p> <p>c. Artigo antes de um pronome possessivo</p> <p>d. Sujeito - Pronome</p> <p>e. Preposição</p> <p>4. Coordenação</p> <p>a. SNs SVs</p> <p>b. Orações</p> <p>c. Outra</p>	<p>5. Resolução incorrecta do POS-homógrafo</p> <p>a. -ing N/V (ex: building, dancing, writing)</p> <p>b. -ing N/Adj (ex: interesting, lasting)</p> <p>c. -ing V/Adj (ex: running, loving, working)</p> <p>d. -ing Adj/Ger_Part. Pres. (ex: promising, following)</p> <p>e. -ed/-en Adj/V_Part. Pass. (ex: alleged, employed)</p> <p>f. -ed/-en Adj/V_Pretérito (ex: experienced, broken)</p> <p>g. -ed V_Pret./V_Part. Pass. (ex: employed, transmitted)</p> <p>h. Adv/Prep/Subordinador (ex: before)</p> <p>i. Adv/Subordinador (ex: as)</p> <p>j. Adj/Adv (ex: early)</p> <p>k. N/V (ex: fight)</p> <p>l. V/Adj (ex: narrow)</p> <p>m. N/Adj (ex: red)</p> <p>n. N/V/Adj/Adv/Prep (ex: round)</p> <p>o. N/Adj/Adv (ex: weekly)</p> <p>p. Prep/V (ex: following)</p> <p>q. Prep/Adj (ex: opposite)</p> <p>r. Prep/Adv/Adj (ex: outside)</p> <p>s. Prep/Subordinador (ex: than)</p> <p>t. Outra</p>

Fonte: Sarmiento (2007 apud Ribeiro 2006, p. 61).

As categorias utilizadas para os fins do presente estudo se limitam especificamente às escolhas léxico-semânticas presentes na tabela de problemas morfológicos e léxicos. Tais classificações são necessárias para identificar os erros de tradução que potencialmente se relacionam com parte da imagem a que se referem.

É importante salientar que o presente estudo se baseia em tais categorias que visam a avaliação de tradução automática numa perspectiva *black box* (portanto, objetivam avaliar a qualidade da mesma) para informar sua análise exploratória. Contudo, não se pretende, conforme delimitado na introdução, qualificar e avaliar efetivamente as relações entre texto e imagem presentes em documentos multimodais traduzidos automaticamente. Tampouco se busca comparar os achados com traduções humanas, localização de software ou estudos de usuário. O que se busca aqui tão somente é identificar e categorizar as potenciais relações que se constituem em documentos multimodais a partir de erros de tradução encontrados neles.

Outros autores da literatura de LC tais como Dorr (1994) trabalham com outros aspectos acerca dos resultados gerados pelas TAs. A pesquisadora (DORR, 1994) apresenta uma descrição formal e propõe uma solução para o problema de classificação de *translation divergences*. Segundo Dorr (ibid., p. 599, tradução nossa), a chamada divergência de tradução se caracteriza quando a mesma informação está representada em ambas as frases da língua-fonte e da língua-alvo, mas a estrutura das frases é diferente.

Dorr (1994, p. 598) cataloga alguns tipos de divergências de tradução conforme a figura abaixo:

é representada como verbo principal em inglês, porém torna-se um modificador adverbial (*gern*) em alemão. O quarto tipo de divergência é a "estrutural", na qual o objeto verbal é representado como uma locução substantiva (*the house*) em inglês, e como uma locução prepositiva (*en la casa*) em espanhol. O quinto tipo de divergência é a "conflacional", ou seja, onde há incorporação de participantes necessários para a realização de uma ação. Nesse exemplo, a frase em inglês utiliza uma única palavra *stab* para as duas palavras em espanhol *dar* (*give*) e *puñaladas* (*knife-wounds*). Este fenômeno acontece devido ao efeito da ação (isto é a porção de símbolo lexical³⁵ de *knife-wounds*) é conflacionada no verbo principal em inglês. O sexto tipo é "categorial", onde o predicado é adjetivado (*hungry*) em inglês, e nominalizado (*Hunger*) em alemão. A última categoria descrita por Dorr (1994, p. 599) é a "lexical", na qual o evento é representado lexicalmente como o verbo principal *break* em inglês, porém como um verbo diferente, *forzar*, (*force* em inglês) em espanhol.

Kameyama et. al. (1991, p. 194) focaliza outros aspectos da classificação de erros de tradução. Os autores definem o conceito de *translation mismatches*, ou "incompatibilidades de tradução", a partir da identificação de situações em que a "gramática de determinada língua não faz uma distinção requisitada pela gramática de da outra língua", destacando características como "a informação específica sobre a

³⁵ Tradução de *lexical token*.

definitude³⁶ e número de determinados substantivos contáveis (ex. *a town, the town, towns, the towns*). Para Dorr (1994, p. 599), o escopo do conceito de "incompatibilidade de tradução" está na informação divergente na relação entre texto-fonte e texto-alvo.

Kameyama et. al (1991, p. 194) destacam duas importantes consequências na tradução automática para grandes incompatibilidades entre as línguas referindo-se a informações contextuais, conforme a passagem a seguir:

A primeira [consequência] em traduzir uma frase da língua-fonte é que as incompatibilidades podem forçar um indivíduo a se basear em informação que não está expressa na frase – informações apenas inferenciáveis a partir do contexto, na melhor das hipóteses. E a segunda [consequência] é que as incompatibilidades podem necessitar da explicitação das informações que estão somente implícitas na língua-fonte ou seu contexto.

Portanto, é a partir do conceito de *translation mismatches* (KAMEYAMA et. al., 1991) que o presente estudo busca identificar as novas relações intersemióticas geradas a partir de tradução automática.

A próxima subseção abordará alguns aspectos teóricos da multimodalidade, os quais o presente estudo se propõe a relacionar com os conceitos abordados nesta subseção, em especial as incompatibilidades de tradução e as categorias (tipologia) de erros de tradução automática.

³⁶ Tradução nossa do termo em inglês *definiteness*.

3.2 Relações texto-imagem

Esta subseção busca abordar o segundo grupo de bases teóricas que informam a investigação do presente estudo. A proposta de relação dos estudos de classificação de erros de tradução automática e da multimodalidade se dá no sentido em que a primeira trata exclusivamente do processamento do componente textual ao qual a segunda, aquela da multimodalidade, se interliga ao componente visual em um mesmo documento. Portanto, desenvolve-se aqui os conceitos e estudos que fundamentam as relações entre texto e imagem dentro do escopo desta pesquisa.

Os estudos abordados aqui sobre as relações texto-imagem estão dentro do tema geral da área abordada na seção 2.6 multimodalidade. Bateman (2014) oferece à perspectiva da multimodalidade o aspecto da multiplicação de sentido por meio da forma como os modos visual e textual são combinados. Bateman (2014, p. 8-10) questiona a visão “natural” de que texto e imagem são dois componentes completamente distintos, colocando exemplos como a representação de um composto orgânico e um mapa.

Esse é o ponto de partida que Bateman (2014) toma para demonstrar diversos problemas e abordagens elaboradas dentro da literatura da multimodalidade. Entre

essas abordagens podemos destacar a “textura intersemiótica³⁷” (Ibid., p.171) dentro da unidade de “coesão multimodal e relações texto-imagem” acomodadas no módulo de “abordagens baseadas no sistema linguístico”.

Para o autor, (BATEMAN, 2014) o trabalho mais significativo dentro dos aspectos de textura intersemiótica é aquele de Liu e O’Halloran (2009). Segundo Bateman (ibid., p. 171), a substancialidade do trabalho dos autores se dá pela expansão dos conceitos de complementaridade intersemiótica de Royce (1998, 2007) e da semiose intermodal no discurso matemático de O’Halloran (2005)³⁸, objetivando oferecer mais um modelo para a integração de diferentes modalidades, do que apenas a documentação de *links*.

O trabalho de Liu e O’Halloran (2009, p. 367) intitulado *Intersemiotic texture: analyzing cohesive devices between language* propõe a “textura intersemiótica” como propriedade crucial de textos multimodais coerentes e apresenta um modelo preliminar para mecanismos coesivos entre língua e imagens” (BATEMAN, 2014). Baseados na proposta de Halliday e Hasan (1976, p. 1-2) de que “textura” envolve relações de significados e constituem elementos cruciais de um texto linguístico, Liu e O’Halloran

³⁷ Abordagem baseada em “modelos de coesão enquanto um recurso semiótico acomodado na semiótica social sistêmico-funcional” - linha que aborda fenômenos que relacionam aspectos da linguística sistêmico-funcional, teoria sobre estrutura linguística baseada em Halliday (1976, 1985, 1994) e Halliday e Matthiessen (2004), e da semiótica social, uma teoria sobre linguagem na sociedade, baseada em Halliday (1978).

³⁸ Ambos conceitos de “complementaridade intersemiótica” (ROYCE, 1998; 2007) e de “semiose intermodal” (O’HALLORAN, 2005) não são explorados no presente estudo.

(2009, p. 369) adicionam o termo "intersemiótico" para tratar das relações semânticas entre texto e imagem representadas por elementos coesivos intersemióticos em um discurso multimodal.

O conceito de "multimodalidade" trabalhado no presente estudo está situado em seu sentido geral, referindo-se à área de investigação, ao invés da pesquisa da relação entre diferentes modos de comunicação; e por isso, o presente estudo se assemelha à distinção utilizada em Liu e O'Halloran (2009) ao dar preferência ao uso de "multisemiótico" como termo de trabalho.

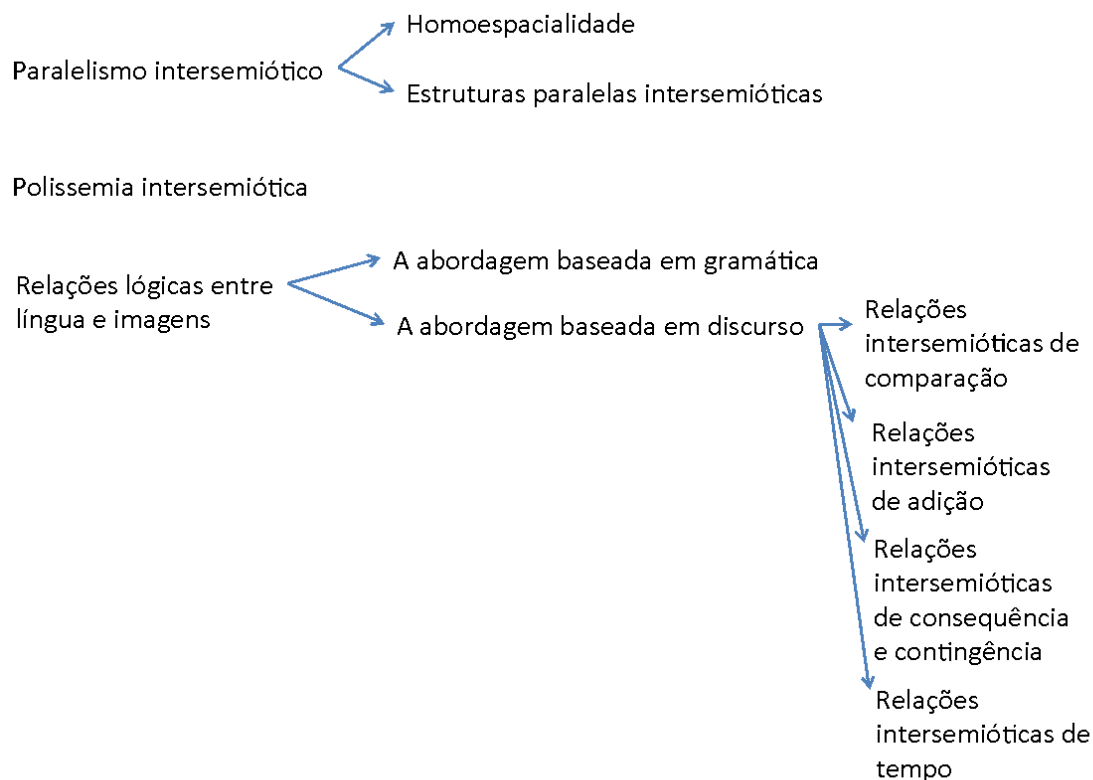
De acordo com O'Halloran (2005, p. 20-21, tradução nossa) "multisemióticos" são os textos que utilizam mais de um recurso semiótico, ou seja, que utilizam mais de um meio para realizar significado; e "multimodal" é utilizado para os textos que envolvem mais de um canal de semiose³⁹ (por exemplo visual, auditivo, e somático⁴⁰).

Assim, os autores apresentam uma tentativa preliminar de categorizar a textura intersemiótica em texto multisemiótico conforme a figura abaixo:

Figura 4 - Categorias de textura intersemiótica propostas por Liu e O'Halloran (2009)

³⁹ O termo utilizado refere-se aos "atos de significado realizados por meio de escolhas a partir da língua e outros sistemas de signos" (O'HALLORAN, 2005, p. 3, tradução nossa).

⁴⁰ Modo relativo ao corpo e/ou parte dele.



Fonte: Próprio autor.

O modelo proposto por Liu e O'Halloran (2009, p. 372-374) é composto por três grandes categorias chamadas “paralelismo intersemiótico”, “polissemia intersemiótica”, e “relações lógicas entre língua e imagens”. A primeira delas, “paralelismo intersemiótico”, se dá por meio de uma relação coesiva que interconecta língua e imagens quando ambos componentes semióticos dividem uma forma similar. Esse paralelismo se constitui por meio da “homoespacialidade” ou pelas “estruturas paralelas intersemióticas”. A primeira se caracteriza pelo paralelismo entre língua e imagens no plano de expressão. E a segunda, se caracteriza quando língua e imagem

compartilham uma configuração de "transitividade"⁴¹ parecida. Para ilustrar esta subcategoria os autores utilizam a figura abaixo:

Figura 5 - Exemplo de estrutura paralela intersemiótica



Israeli army dog attacks Palestinian woman

Fonte: Liu e O'Halloran (2009, p. 373).

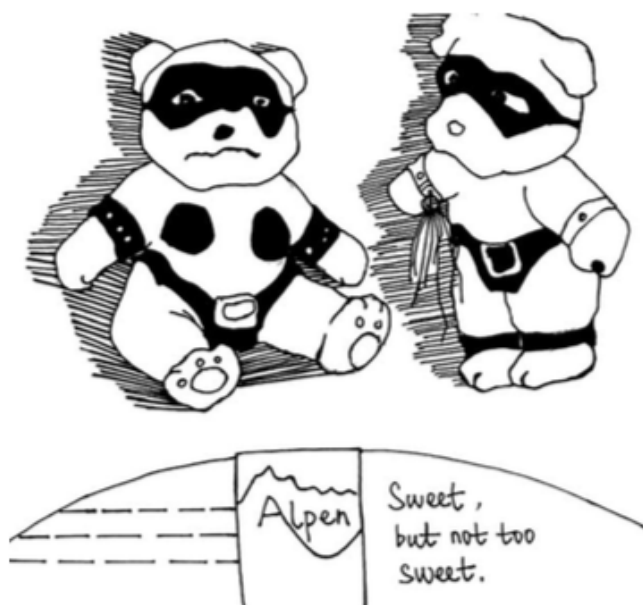
Segundo Liu e O'Halloran (2009, p. 373-374), a figura acima apresenta uma ação de um cachorro mordendo uma mulher muçulmana. Essa ação, representada por um processo material, físico, (dentro do sistema da gramática da transitividade) é também compartilhado na descrição da legenda "Israeli army dog attacks Palestinian woman"

⁴¹ Sistema gramatical composto por tipos de processos (descrito pelo verbo), participantes do processo, e circunstâncias relacionadas ao processo (para mais informação vide Halliday e Matthiessen, 1985, 1994, e 2004).

(cão do exército israelense ataca mulher palestina). Essa relação, portanto, pode ser caracterizada como uma estrutura paralela intersemiótica.

Outra categoria descrita por Liu e O'Halloran (2009, p. 375) é a “polissemia intersemiótica”. Nela, a relação coesiva entre os componentes verbais e visuais que compartilham significados múltiplos em textos multisemióticos. Essa categoria compartilha significados similares em oposição a significados diferentes, gerando o que os autores chamam de “relações de co-contextualização” (ibid. 375). Liu e O'Halloran (2009, p. 375) utilizam a seguinte ilustração para descrever esta categoria:

Figura 6 - Exemplo de polissemia intersemiótica



Fonte: Liu e O'Halloran (2009, p. 375).

A figura acima apresenta uma publicidade de cereais *Alpen*. Nela, a relação de significados entre o anúncio *Sweet, but not too sweet*, (em português “doce, mas não

muito doce”) em contraponto com a leitura de dois ursos de pelúcia utilizando acessórios sadomasoquistas, gera um resultado polissêmico (LIU e O’HALLORAN, 2009, p. 376-377). Além disso, para os autores (Ibid., p. 375) tal relação texto-imagem utiliza-se de outras relações intersemióticas que cooperam com a polissemia intersemiótica, tais como a “eclipse intersemiótica” (LEMKE, 1998) e a “correspondência intersemiótica” (JONES, 2006).

A correspondência intersemiótica, diferentemente de um sinônimo ou repetição, caracteriza a relação entre um elemento linguístico e outro visual com vistas para o uso conjugado entre significados verbais e visuais visando uma correspondência e a expansão do significado resultante (JONES, 2006, p. 194).

Já a eclipse intersemiótica (O’HALLORAN, 2005 baseado em LEMKE, 1998) se realiza quando a imagem ou parte dela é criada para compensar a falta de construções gramaticais como, por exemplo, o recurso de tabela enquanto “apresentação visual textualizada” (LEMKE, 1998).

Em relação à terceira e maior categoria de textura intersemiótica, Liu e O’Halloran (2009) buscam tratar das relações lógicas entre língua e imagens, isto é da análise de significados lógicos entre componentes verbais e visuais baseada em dois tipos de abordagem: (i) gramatical e (ii) discursiva.

Segundo Liu e O’Halloran (2009, p. 377) a abordagem de relações lógicas entre

língua e imagens baseada em gramática⁴²:

[...] fornece uma descrição preliminar do significado lógico entre diferentes recursos semióticos em mídias novas e mídias antigas, nas quais língua e imagem são consideradas iguais ou diferentes entre si em termos de *status* relativo enquanto as relações intersemióticas lógico-semânticas de Projeção e Expansão⁴³ se aplicarem (LIU e O'HALLORAN, 2009, p. 377, tradução nossa).

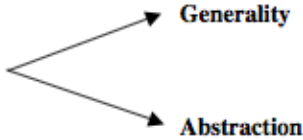
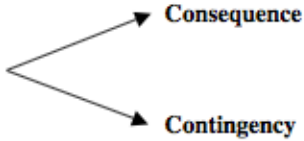
Nessa passagem os autores deixam claro a limitação analítica em torno do fenômeno das relações lógico-semânticas entre língua e imagens, e diante disso apresentam a necessidade de se expandir e complementar tais relações baseadas em gramática com o nível discursivo (baseado em O'Halloran 2005).

As quatro subcategorias de relações intersemióticas da abordagem discursiva, a saber “comparativa”, “aditiva”, “consequencial”, e “temporal”, podem ser visualizadas a partir da tabela a seguir proposta por Liu e O'Halloran (2009, p. 384):

Tabela 8 – Relações intersemióticas lógicas

⁴² Abordagem gramatical baseada em Martinec e Salway (2005).

⁴³ De acordo com Halliday, (HALLIDAY, 1985) há dois tipos principais de relações lógico-semânticas entre orações: a de projeção (envolvendo locuções e ideias) e a de expansão (envolvendo extensão, elaboração, e encarecimento – *enhancement* em inglês).

Logical Relations	Meaning
Comparative 	Similarity
Additive	Addition
Consequential 	Cause Purpose
Temporal	Successive

Fonte: Liu e O'Halloran (2009, p. 384)

A tabela acima ilustra os 4 tipos de relações lógicas baseadas no discurso e seus respectivos significados. De acordo com Liu e O'Halloran (2009, p. 379) a relação intersemiótica de "comparação" é um tipo de recurso utilizado para organizar o significado lógico em relação a similaridade entre os componentes linguísticos e visuais no discurso multimodal, reformulando semioticamente um ao outro. Tais reformulações se realizam no nível de "generalidade" (por exemplo, quando a relação lógico-semântica entre o componente linguístico e visual se realiza por meio da relação geral-específico), e de "abstração" (nos casos onde a reformulação lógico-semântica entre parte do componente visual e linguístico se concretiza a relação concreto-abstrato). Já a subcategoria de "adição" se caracteriza quando um componente semiótico adiciona uma

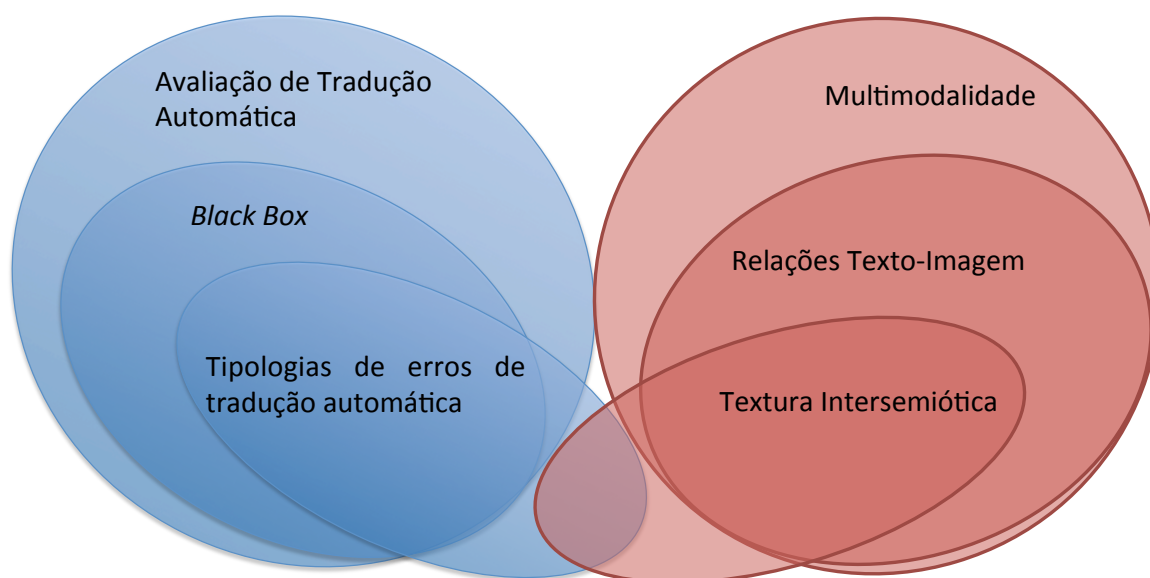
nova informação ao outro componente semiótico. A terceira categoria, "consequencial" pode ser identificada quando uma mensagem semiótica é percebida "possibilitando" ou "determinando" a outra mensagem ao invés de simplesmente precede-la (MARTIN, 1992, p. 193 apud LIU e O'HALLORAN, 2009, p. 380). De acordo com os autores (ibid. p. 380) as relações intersemióticas de consequência podem ser sub-classificadas como "consequência" e "contingência", nas quais a primeira refere-se a relações causais não-modalizadas entre mensagens verbais e visuais onde o efeito foi assegurado; já a segunda se caracteriza em textos multisemióticos onde a causa tem apenas o potencial para determinar a possibilidade, porém sem haver efeito assegurado. Por fim, a subcategoria de relação intersemiótica "temporal" se realiza quando os passos (sequências) de um procedimento são representados verbalmente e visualmente (geralmente encontrados em textos como manuais e guias ilustrados) (IEDEMA, 2003, apud. LIU e O'HALLORAN, 2009, p. 383).

Assim, o modelo preliminar de Liu e O'Halloran (2009) acerca da concepção de "texturas intersemióticas" contribui amplamente para o presente estudo. Ele oferece categorias fundamentadas na literatura da abordagem da multimodalidade, mais especificamente na linha da semiótica social sistêmico funcional, e propõe a expansão da abordagem gramatical para a discursiva.

3.3 Considerações finais

Abordou-se no presente capítulo os subsídios teóricos das categorias de erros de tradução automática, e das relações texto-imagem que informam a análise do presente estudo. Forma-se, portanto, uma interface mais específica que pode ser representada da seguinte forma:

Figura 7 - Interface teórico-analítica



Fonte: Próprio autor.

Como pode-se notar na figura acima, as subáreas de cada arcabouço teórico se encontram na intersecção onde se delimita o problema sob investigação, qual seja, aquele das relações texto-imagem formadas a partir da tradução automática de

documentos multimodais (especificamente, *webpages*, manuais, infográficos, e artigos acadêmicos).

O capítulo seguinte trata dos procedimentos metodológicos elaborados para a investigação do problema de pesquisa.

4. Métodos

O presente capítulo tem por finalidade descrever os métodos e instrumentos utilizados para responder às seguintes perguntas de pesquisa:

1. Como a abordagem da multimodalidade contribui para a avaliação de resultados de tradução automática em documentos multimodais?

2. Como se dão as relações texto-imagem em documentos multimodais a partir dos resultados da tradução automática?

Observa-se a partir das perguntas de pesquisa acima, que a presente investigação se caracteriza por método exploratório e qualitativo. Esse método de pesquisa foi identificado a partir dos achados que emergiram do estudo piloto (explicado na seção 4.2). Então faz-se necessário abordar, ainda que brevemente o traçado metodológico que levou o estudo até o ponto onde foi necessário reconfigurar o escopo teórico-metodológico. Assim, a primeira parte dos métodos trata brevemente (já que os dados demonstraram que não estavam no cerne do problema) dos "procedimentos metodológicos" (seção 4.1). A seguir, detalha-se o "estudo piloto" (seção 4.2); e por fim discute-se os resultados (seção 4.3) obtidos do estudo piloto, os quais apontam para a posterior etapa de análise da base de dados da pesquisa, descrita no capítulo 5.

4.1 Procedimentos metodológicos

Inicialmente, o percurso investigativo se propunha a “identificar se a anotação multimodal de um corpus de relação discursiva multimodal contribui para a tarefa de avaliação de tradução automática”. O objetivo era o de identificar as relações texto-imagem a partir do paradigma da relação retórica multimodal, baseada no arcabouço teórico-metodológico de Bateman (2008) e, mais especificamente de Hiippala (2013), e Taboada e Habel (2013).

Como a proposta de pesquisa apresenta um problema específico dentro da interface de pesquisa dos estudos de TA, multimodalidade, e CI, parte da estrutura conceitual dessas áreas deve informar a amostra para analisar se há relações retóricas (coesão entre texto gerado pela tradução automática em relação a outros modos de representação visual no artefato multimodal⁴⁴, tais como fotografias, infográficos, e tabelas), e se houver tais relações, quais são, e como elas se caracterizam e operam para criar um sentido unificado; ou se não há relações retóricas identificáveis.

As relações retóricas em questão advém da "teoria descritiva sobre a organização de textos naturais, que se baseia em argumentos principais, quais sejam (i) do predomínio de padrões estruturais de núcleo/satélite, (ii) a base funcional

⁴⁴ De acordo com Bateman (2008), um exemplo de artefatos multimodais é um documento multimodal, no qual "...uma variedade de modos baseados no visual organizam-se simultaneamente para preencher uma coleção orquestrada de objetivos comunicativos interligados entre si." (tradução livre, p. 1).

hierárquica, e (iii) o papel comunicativo da estrutura do texto" de Mann e Thompson (1988, p. 243), mais conhecida como *Rhetorical Structure Theory* (RST).

Contudo, de acordo com os achados do estudo piloto (vide seção 4.2), o que se nota a partir da tentativa de identificação das relações retóricas entre texto e imagem em *webpages* é que a organização retórica multimodal não se modifica na tradução, já que a tradução automática i) não é desenvolvida para reconhecer padrões de ligações coesivas entre modo textual e visual para informar suas traduções (como acontece frequentemente em traduções humanas intersemióticas, tais como nos gêneros de quadrinhos⁴⁵, tradução audiovisual e localização⁴⁶; e ii) que tais modificações emergem a partir de modificações de sentido dentro do modo textual, ocasionadas pelo resultado de tradução automática, ou seja, de uma incompatibilidade (*mismatch*) dentro da estrutura textual que se refere semanticamente a um componente visual, ou parte dele.

A partir de tais achados, entendeu-se que o percurso metodológico é exploratório, pois requer um *design* que pode ser testado com um estudo piloto antes de coletar quantidades mais significativas de dados para verificar se o padrão é representativo de anotação de uma base de dados para análise de geração de tradução por TA.

⁴⁵ Para maiores informações sobre multimodalidade e tradução de história em quadrinhos vide Borodo (2015).

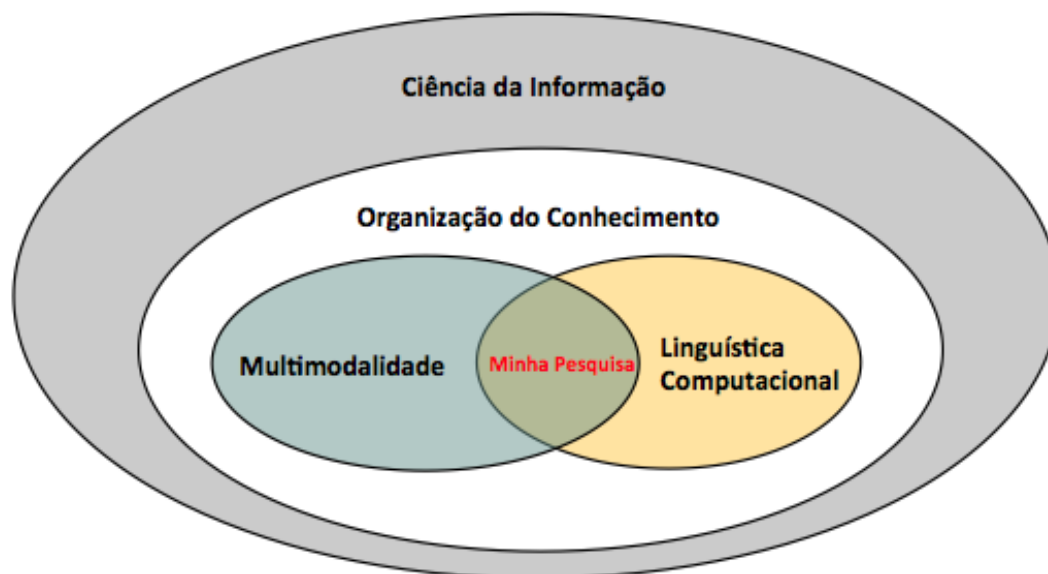
⁴⁶ Para maiores informações sobre a relação entre textos multimodais e a indústria de multimídia vide Ricoy (2012).

De acordo com Creswell (2009, p. 11), uma visão de pesquisa pragmática “observa o *o quê* e o *como* da pesquisa, baseado nas consequências desejadas de onde se queira levar a cabo”. O autor (ibid.) afirma que uma perspectiva de mundo mais pragmática favorece uma estratégia de pesquisa com métodos mistos. Portanto, esta investigação adota a estratégia de métodos mistos, pois busca por padrões a partir do entendimento das seguintes perguntas de pesquisa: i) "como a abordagem da multimodalidade contribui para a avaliação de resultados de tradução automática em documentos multimodais?"; e ii) "como se dão as relações texto-imagem em documentos multimodais a partir dos resultados da tradução automática?". Dentro dessa perspectiva, o presente estudo adota especificamente uma estratégia exploratória sequencial, que envolve uma "primeira fase onde uma teoria ou conceito é testado, seguida por um método qualitativo envolvendo uma exploração detalhada com alguns casos e indivíduos" (CRESWELL, 2009, p.14).

Todavia, a razão por esta tese priorizar uma abordagem mais qualitativa (exploratória) se deve ao fato de valores significativos para análise ainda serem escassos na literatura estudada, e mais especificamente dentro do contexto da interface interdisciplinar descrita.

Assim esta tese explora a seguinte interface interdisciplinar como arcabouço teórico e metodológico.

Figura 8 - Interface interdisciplinar



Fonte: Próprio autor.

Dentro dessa estrutura interdisciplinar, a presente proposta metodológica se propõe a responder as perguntas de pesquisa por meio da utilização dos aportes teóricos estudados no capítulo 3 (Marco Teórico), quais sejam, aqueles das categorias de erros de tradução automática (baseado em VILAR et. al., 2006) e do conceito de incompatibilidade de tradução (baseado em KAMEYAMA et. al., 1991), localizados dentro da área da Linguística Computacional; e do modelo de textura intersemiótica proposto por Liu e O'Halloran (2009), localizado na área da Multimodalidade.

A seção 4.2. traça os procedimentos utilizados para se alcançar o objetivo deste estudo, qual seja, o de "identificar relações semânticas entre texto e imagem em

documentos multimodais traduzidos automaticamente para a avaliação de tradução automática". A seção 4.3 redefine alguns procedimentos em função dos achados do estudo piloto. A próxima subseção, 4.1.1 trata do instrumental de pesquisa e os critérios utilizados para sua escolha.

4.1.1 Instrumentos e materiais

O percurso metodológico utiliza *webpages* como base de dados para análise. A justificativa para tal escolha se deve mais precisamente à busca em documentos hipertextuais conectados na *World Wide Web* conteúdos visuais e verbais utilizados para criar uma unidade semântica, tais como fotografia com legendas e títulos referentes a imagens dentro de artigos de conteúdo jornalístico, educativo e procedimental.

Como o problema de pesquisa da qualidade de tradução está contida dentro do problema de informação multimodal em tradução gerada automaticamente, esta pesquisa entende que o *Google Translate* representa um amplo meio de traduzir automaticamente conteúdo de *websites*. Assim, a escolha deste padrão é justificada como instrumento para traduzir o conteúdo dos websites do inglês para o português. O critério de seleção da língua estrangeira se deve às habilidades linguísticas do analista.

Os dados incluem *websites* de conteúdo multimodal tais como manuais ilustrados, enciclopédias, portais de notícias e websites comerciais. A escolha desses tipos de conteúdos se justifica pelo potencial de apresentar recursos semióticos como fotografias, desenhos, quadros, legendas, e manchetes. Tais recursos foram identificados na literatura da multimodalidade e mais especificamente no estudo piloto desta pesquisa como elementos que podem apresentar novas configurações intersemióticas quando se utiliza o Google Tradutor. Tais recursos semióticos devem ser identificados no nível de suas relações semânticas em inglês e comparados com a tradução gerada pela TA em português.

O *Google Translate*, ou Google Tradutor⁴⁷ na versão em português, é um serviço de tradução automática multilíngue e online, oferecido gratuitamente ou pago (no caso de pessoas físicas e jurídicas que queiram implementar o serviço no desenvolvimento de um aplicativo ou website). Por esse motivo, e pelo método de ATA chamado *black box* (vide seção 2.3) utiliza-se o serviço oferecido gratuitamente⁴⁸.

Esta subseção descreve brevemente alguns métodos, instrumentos e materiais utilizados para o momento da investigação que leva até o estudo piloto. A próxima seção elabora o estudo piloto, que descreve em detalhes a aplicação dos procedimentos e

⁴⁷ Para maiores informações sobre o Google Tradutor, vide a seção 2.3 esta tese.

⁴⁸ A pesquisa entrou em contato com a representação dos serviços do Google Tradutor, informando sobre os objetivos da mesma. Os documentos sobre essa conversa encontram-se nos anexos da tese.

instrumentos projetados até aqui.

4.2 Estudo piloto

A razão para desenvolvimento do estudo piloto é metodológica. Como há uma escassez de estudos que investigam a compilação de uma base de dados específica, é importante adotar uma estratégia prudente por meio de teste da metodologia com uma amostra menor.

O estudo piloto ⁴⁹ iniciou com a seleção de *websites* com conteúdos representando informações visuais e verbais de língua inglesa. Examinou *websites* como *National Geographic*, *National Geographic Kids*, *BBC news*, *BuzzFeed*, *Wikihow*, *The Economist*, *Science Daily*, *Recipe blogs*, *Amazon*, *Deal Extreme*. O motivo de ter escolhido esses *websites* se pauta pela presença de artigos escritos com componentes linguísticos e visuais em relação semântica, tais como legenda e fotografia, e subtítulos referindo-se a desenhos, por exemplo. Tais componentes podem ser encontrados na literatura da multimodalidade, e mais especificamente nos estudos de relações intersemióticas. A quantificação dessa seleção não foi considerada, pois este estágio exige entendimento das relações entre imagens e textos na tradução automática.

⁴⁹ A realização do estudo piloto se deu durante estágio de doutorado-sanduíche da CAPES no *Bremer Institut für Transmediale Textualität*, na *Universität Bremen*, de abril de 2015 a dezembro de 2015, sob supervisão do prof. dr. John Bateman. Segue em anexo parecer do professor supervisor, atestando as atividades realizadas durante o estágio.

A primeira fase consistiu na tradução de *webpages* automaticamente do inglês para o português por meio da ferramenta Google Tradutor disponível no navegador *Google Chrome*. A escolha pelo navegador *Google Chrome* se deu pela compatibilidade com o Google Tradutor, já que as duas ferramentas pertencem à mesma empresa. A seguir, ambas as versões foram capturadas como imagem por meio do *add-on FireShot* e salvas no computador para posterior análise manual. A escolha da ferramenta de captura de imagens *Fireshot* se justifica pela gratuidade, capacidade de escanear o conteúdo da página *web* completo, e de poder salvar as capturas de imagem em diversos arquivos de imagem ou em pdf, disponíveis para transferência para o computador de seu usuário.

O seguinte procedimento foi realizado para a preparação da base de dados para análise:

- Passo 1 - Organização dos dados por pastas de gênero e subgênero textuais⁵⁰;
- Passo 2 - Organização dos dados de amostra por gêneros e subgêneros bilíngues;
- Passo 3 - Alinhamento de ambas as versões original e tradução de todas as páginas transcritas manualmente com suas traduções automáticas online;
- Passo 4 - Identificação de exemplos de relações visuais-verbais baseadas em RST.

Utilizou-se programas como o *Preview* para mac, *Sketch*, e *UAM tools* para anotar

⁵⁰ Levou-se em consideração a noção informal de gênero textual tal qual proposta por Bateman (2008).

qualquer relação semântica reconhecida entre texto e imagem. A escolha das duas primeiras ferramentas se deve ao fato da possibilidade de uma anotação manual e com boa visualização dos elementos identificados em arquivos .png, .jpg, e pdf, utilizados nesta pesquisa. Já a escolha do UAM tools, se deve ao seu potencial de anotar esquemas de parte ou todo de uma imagem, e ao final quantificar os tipos de categorias analisadas.

Nesse processo pode-se observar que as páginas *web* são frequentemente fixas para variação de discurso multimodal, que é uma variação entre os modos que criam um sentido unificado entre aqueles criados por folder publicitário, quadrinhos, filmes e outros artefatos. De fato, imagens de uma *webpage* não mudam quando o texto é automaticamente traduzido.

Essa observação segue parcialmente o padrão geral de se manter o aspecto visual na tradução de quadrinhos, mas difere em relação ao aspecto verbal na tradução. A razão para essa diferença se dá pelo fato de a TA não utilizar significado visual (de imagem; não confundir com *OCR*) como um recurso de tradução, tal qual um tradutor humano o faria (BORODO, 2015). A TA reproduz uma relação estatística e probabilística na sua base de dados. Assim, se parte da *webpage* que relaciona semanticamente texto e imagem é automaticamente traduzida, logo, parte do significado multimodal do texto original "modifica-se".

Dessa forma, uma imagem no documento original é a mesma imagem na

tradução automática online de uma *webpage*, e ocupa o mesmo espaço na página; o texto também ocupa o mesmo espaço, porém com certas “incompatibilidades” (KAMEYAMA et. al., 1991) a partir dos resultados gerados pela TA. Em outras palavras, pode haver mudança de significado na tradução, e, portanto, mudanças de significado em relação a outros modos que se combinam com o componente linguístico, criando uma nova configuração de sentido no documento traduzido automaticamente.

Os achados mostram que a maior parte das páginas representam pouca mudança na relação de significado visual-verbal como um todo, devido ao alto nível de traduções corretas lexico-semânticas que podem ocorrer no texto. Quando uma ou outra palavra é traduzida, o restante do texto completa essa lacuna semântica, ou até mesmo o elemento visual em si desempenha esse papel (quando diretamente referindo-se ao texto).

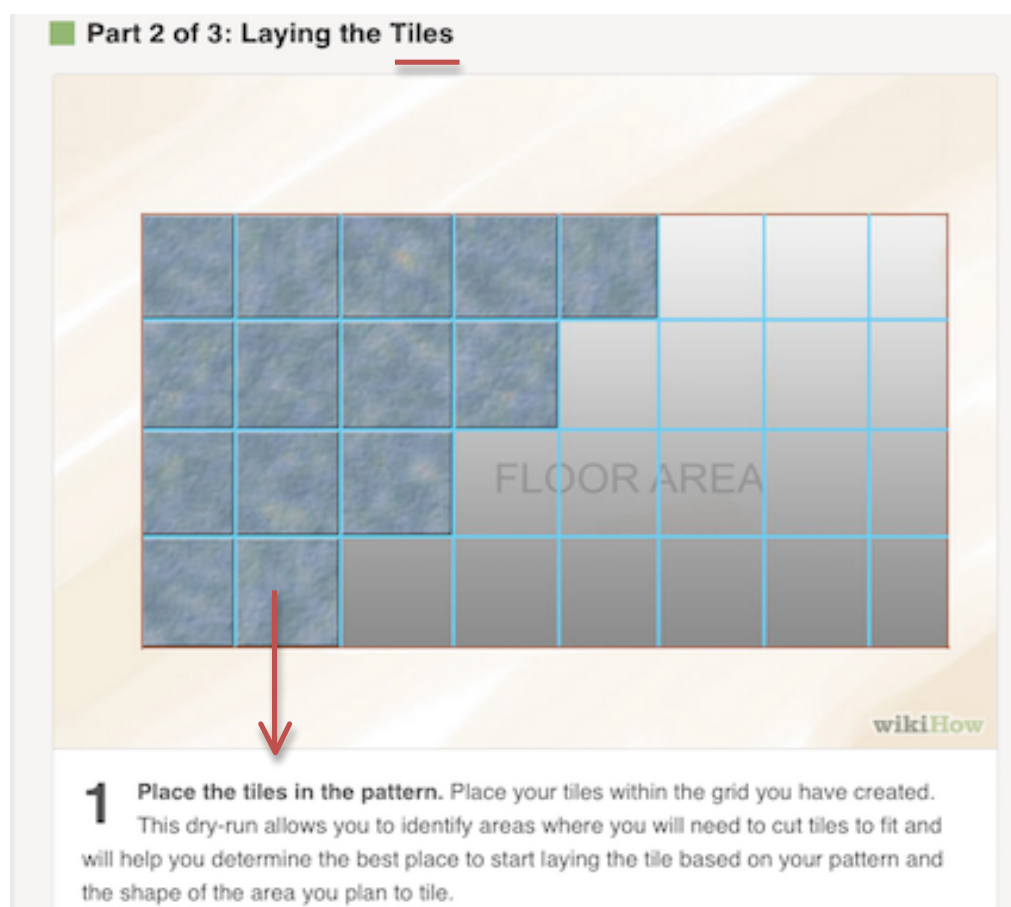
A única forma pela qual se pode reconhecer essas “mudanças” de significado visual-verbal na tradução é quando há incompatibilidade (*mismatch*) léxico-semântica referindo-se a uma imagem ou parte dela, ou seja, a imagem não representa os componentes semânticos que faltam na tradução. Por essa razão intitulou-se tal fenômeno de “incompatibilidade intersemiótica” (*intersemiotic mismatch*) a partir da geração de texto automaticamente traduzido.

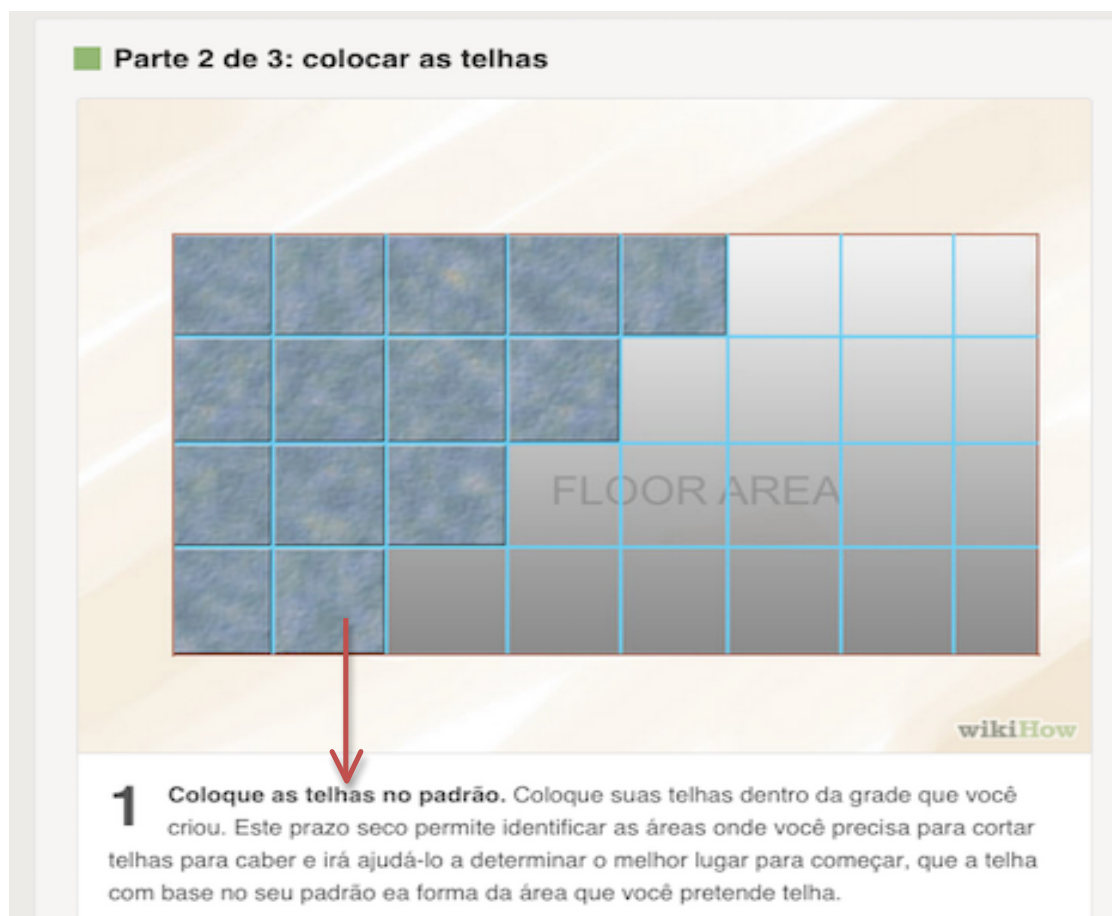
Kameyama et al. (1991) sugerem que as incompatibilidades de traduções

automáticas “encontradas quando a gramática de uma língua não faz uma distinção necessária pela gramática da outra língua”. Baseando-se nessa definição, uma incompatibilidade intersemiótica em TA é encontrada quando, por meio da geração de uma tradução por TA constata-se uma nova configuração semântica entre os modos verbais e visuais em relação aos mesmos modos encontrados no documento original.

As unidades básicas de aspectos verbais, tais como unidades léxico-semânticas em legendas de uma imagem, apresentam divergências que mudam seu significado em relação a uma imagem, conforme demonstrado nos exemplos abaixo:

Figura 9 - Imagens capturadas do site *WikiHow* em inglês e sua tradução do Google Tradutor para o português





Fonte: *Wikihow*.

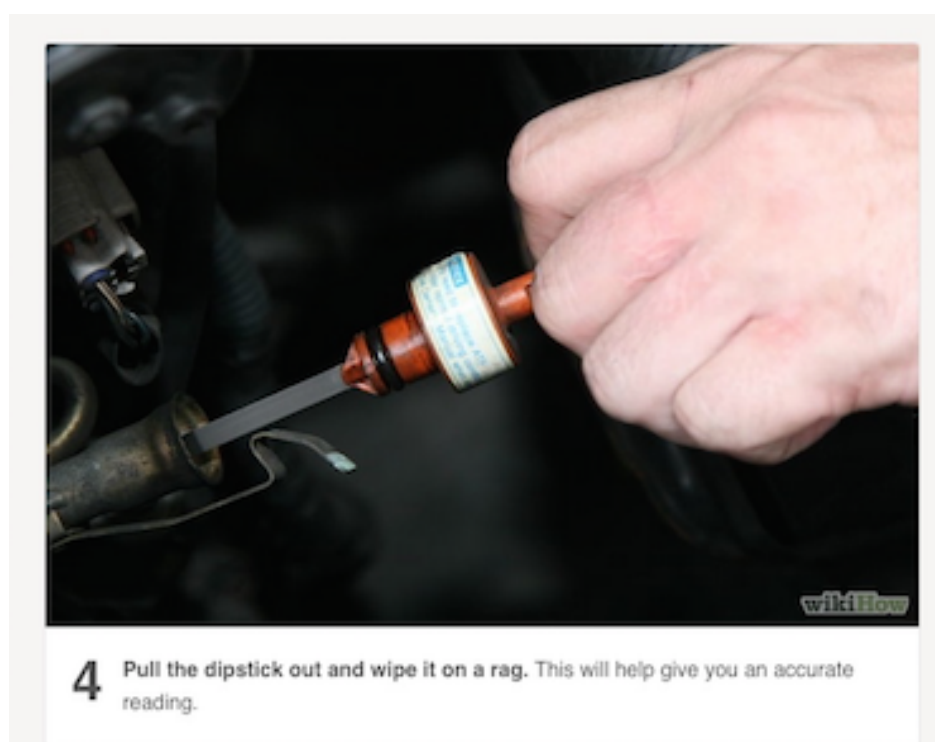
O exemplo acima ilustra um fenômeno envolvendo incompatibilidades intersemióticas léxico-semânticas em textos traduzidos automaticamente. A imagem foi tirada do artigo da *Wikihow* que explica os procedimentos sobre “como instalar piso de mármore”.

Ao examinar a figura acima, observa-se que o item lexical “*tile*” exibe a tradução “telhas”. O uso contextual da palavra “*tiles*” elimina outros possíveis significados para o item lexical. O contexto pode ser expressado pelo contexto visual ou verbal do artigo: mas não é representado visualmente em português. “Telhas” corresponde a “roof tiles”

em inglês, embora a imagem mostre o piso com as palavras "floor area". Neste sentido, o significado visual correspondente de "tiles" representa um significado diferente de "telhas".

Outros dois exemplos mostram essa mudança de significado na página "how to add fluid transmission" na *WikiHow*, e "this woman touching tattoo started a global mental health movement" no *website Buzzfeed*. Os exemplos estão representados na figura a seguir.

Figura 10 - Imagens capturadas do site WikiHow em inglês e sua tradução do Google Tradutor para o português





- 4** Puxe a vareta para fora e limpe-o em um pano. Isso ajudará a dar-lhe uma leitura precisa.



- 5** Stick the dipstick back into the transmission fluid and lift it out again for your reading. You should now be able to see what level your transmission fluid reaches. Remember to read the "hot" level on the transmission dipstick.



Fonte: *Wikihow*

Uma possível tradução em português seria "espete a vareta de volta no fluido de transmissão". O que está representando uma incompatibilidade intersemiótica na imagem é a diferente experiência criada com o sinônimo "furar" junto com a palavra vareta, em relação ao original. Como a ação em inglês é omitida na foto, cria-se uma premissa omitida que precede a terceira ação descrita, "*for your reading*", pois a vareta está visivelmente sobre um pedaço de pano, destacado em relação ao fundo desfocado, que é aparentemente o interior de um carro. Então, a ação omitida na fotografia em relação a uma descrição explícita torna-se uma ação omitida na fotografia em relação a um significado diferente criado pela tradução automática. A consequência de não haver

uma ação explícita no visual do original dá destaque à ambiguidade na relação texto-imagem no caso da tradução.

Abaixo da fotografia em questão há uma legenda com o número do passo, seguido de uma descrição com a primeira frase em negrito.

As relações coesivas de imagem-texto são uma reformulação (MARTIN, 1992; LIU; O'HALLORAN, 2009). O elemento verbal está reformulando o visual. Contudo há partes do visual que são omitidas nessa reformulação, por exemplo "*to stick the dipstick back*" and "*lift it out*". Estas ações não são representadas na figura, mas pode-se sugerir que a vareta mostrada na fotografia já havia sido colocada em seu compartimento e depois levantada, e a fotografia, estaria no estágio de "leitura" em si. Pode-se entender que neste caso a imagem não forma um novo significado juntamente com o componente verbal.

Em relação à tradução automatizada em português, observa-se a mesma incompatibilidade intersemiótica, isto é, especificamente, a ação de "colocar a vareta de volta no fluído de transmissão" ausente na fotografia. Contudo, neste exemplo, há uma imprecisão semântica com a palavra "furar". A imprecisão do tempo verbal na tradução (de imperativo para infinitivo) não é considerada pelo presente estudo, pois o mesmo não conduz relação semântica com o elemento visual analisado.

O exemplo abaixo foi retirado do website *Buzzfeed*. Ele demonstra ambiguidade

(verbo “*drawing*” torna-se “tirando”, que é semanticamente incompatível com o contexto ilustrado, porém representa um significado diferente da narrativa, que é realizado a partir da ação de “tirar a tatuagem”) um novo significado que surge da representação omitida da ação na fotografia.

Figura 11 - Imagens capturadas do site *Buzzfeed* em inglês e sua tradução do Google Tradutor para o português



Fonte: *Buzzfeed*.

A análise dos três exemplos acima baseia-se no conceito de complementaridade intersemiótica de Royce (1998; 2007) para classificar relações semânticas entre os elementos verbais e visuais, e Dorr (1990, 1994) e Kameyama et al. (1991) para as

categorias de incompatibilidade de tradução que informam parte dos fenômenos observados. Contudo, a análise do segundo exemplo, baseia-se também em Liu and O'Halloran (2009) por incluir uma abordagem discursiva à perspectiva gramatical de Royce (2007). Assim, as figuras 9 e 11 são classificadas como “ambiguidade intersemiótica” e a figura 10, como “relação temporal intersemiótica”, respectivamente.

Esta seção descreve o estudo piloto e relata alguns achados. A próxima seção discute os achados e como eles informam uma redefinição parcial no desenho da tese.

4.3 Resultados

Conduziu-se este estudo piloto para identificar possíveis “mudanças” de significado na relação verbal-visual em páginas com conteúdos. A razão de se conduzir este piloto é puramente metodológica, pois não há estudos acerca da compilação de uma base de dados de tal natureza dentro da perspectiva interdisciplinar proposta. Portanto foi uma estratégia prudente ao testar com uma amostra pequena.

O estudo piloto mostrou que a metodologia proposta era parcialmente apropriada para a aplicação em uma base de dados maior. Isto significa que a base linguística da multimodalidade (no que se refere à gramática sistêmico-funcional, por exemplo⁵¹) pode contribuir para a categorização de palavras relacionadas a partes de

⁵¹ vide Halliday and Matthiessen (2004).

imagem em um documento multimodal. Também detectou-se que a RST poderia ser útil em casos como o terceiro exemplo, onde uma relação multimodal baseada em discurso é reconhecida. Então em fases finais de análise (talvez após a fase baseada na gramática) a RST pode ser útil para analisar casos baseados em discurso.

Assim a utilização de Liu and O'Halloran (2009) durante o pré-teste se mostrou satisfatória para a categorização multimodal, e Dorr (1990; 1994) e Kameyama et al. (1991) importantes como base para fundamentar a classificação de incompatibilidades intersemióticas a partir de uma tradução gerada automaticamente. Além disso, a abordagem de Bateman (2014) sobre a relação texto e imagem informam uma análise mais compatível para se entender o problema de pesquisa, do que Bateman (2008).

Em relação aos tipos de documentos multimodais, nota-se que os *websites* são significativamente fixos para variação semântica entre modos. Contudo, algumas páginas *web* mostraram alguns resultados significativos quando automaticamente traduzidos para o português, devido especialmente a presença de legendas próximas à imagem normalmente na forma de infográficos, mapas, manchetes e legendas em sites de notícias tais como *BBCnews* e *Times*, e em sites de "instruções" tais como *Wikihow* e *blogs* de receitas.

Constatou-se que a detecção de discrepâncias na relação visual-verbal em páginas *web* traduzidas automaticamente é uma tarefa que consome muito tempo.

Assim, para a análise dos dados do estudo considera-se apenas os gêneros e subgêneros onde encontrou-se dados para análise. Assim, há mais chances de encontrar amostras para análise. Por essa razão é necessário a realização de uma pesquisa exploratória.

Outro ponto de discussão diz respeito a necessidade de categorização dessas amostras para posterior processamento. Como este estudo exploratório busca “mapear” possíveis características na relação verbal-visual, considera-se não apenas as “incompatibilidades” e “divergências” de tradução automática, mas as “novas configurações intersemióticas” geradas a partir dela.

Assim, a emergência de duas categorias de incompatibilidade intersemiótica evidencia a necessidade de conduzir a pesquisa em direção a uma estrutura teórico-metodológica que permita a análise de novas configurações intersemióticas em documentos multimodais traduzidos automaticamente. A fase exploratória/qualitativa da investigação tornou-se proeminente em relação aos métodos quantitativos/qualitativos sugeridos anteriormente.

Embora a quantificação dos documentos não fosse objetivo desta etapa, chegou-se a um total de 15 pares de documentos multimodais. Portanto, a partir dos achados do estudo piloto, esta pesquisa dobrou este número para a análise principal dos dados. A justificativa se deve ao fato de que é possível obter um número mais amplo e variável dos casos de incompatibilidades intersemióticas encontradas no estudo piloto.

O presente percurso metodológico segue o plano de trabalho encontrado na seção "apêndices" desta tese. Levando em consideração os aspectos levantados neste capítulo e redefinidos com base no estudo piloto, leva-se a cabo o próximo capítulo de análise dos dados.

5. Análise da base de dados

A partir do estudo piloto descrito na subseção anterior, este capítulo explora e expande uma tipologia de erros de TA baseado em Vilar et. al. (2006) , Llitj'os et. al. (2005), no conceito de *translation mismatch* de Kameyama et al. (1991), e no modelo de "textura intersemiótica" proposto por Liu e O'Halloran (2009), como base para uma classificação de erros de tradução em documentos multimodais onde há relações intersemióticas entre texto e imagem .

Este capítulo analisa incompatibilidades intersemióticas (divergências semânticas entre texto e imagem geradas a partir de resultados de erro de TA) em trinta documentos multimodais produzidos em língua inglesa e suas respectivas traduções geradas automaticamente pelo Google Tradutor para a língua portuguesa. A base de dados é composta por alguns tipos de *websites* (notícias, entretenimento, educação), além de manuais ilustrados, artigos acadêmicos e infográficos, com o intuito de ampliar a possibilidade de novas categorias de incompatibilidades intersemióticas, baseado nos achados do estudo piloto.

A preparação dos trinta pares de documentos multimodais seguiram conforme descrito na subseção do estudo piloto para posterior análise manual com base no aparato teórico descrito no capítulo 3. Por questões de restrição na visualização,

decidiu-se exibir apenas alguns exemplos dos dados analisados⁵² neste capítulo. A tabela abaixo demonstra como os dados dos arquivos foram registrados para organização da análise.

Tabela 9 - Informações gerais sobre os documentos analisados

File name	Língua	Fonte	Formato do	Meio	Tipo do conteúdo	Unidades básicas envolvidas
Ways to Build a Bench_wikiHow_pt	Português	http://www.wikihow.com/Build-a-Bench	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos
How to Make Coconut Flour With Flaked_pt	Português	http://www.wikihow.com/Make-Coconut-Flour-With-Flaked-Coconut	.png	Página Web	procedimental/explicativo	fotografias, legendas, subtítulos
Como se livrar das pulgas na	Português	http://www.wikihow.com/Get-Rid-of-Fleas-Naturally	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos
How to Get Into Har_pt	Português	http://www.wikihow.com/Get-Into-Harvard-Business-School	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos
How to Fix a Squeaking Bed_pt	Português	http://www.wikihow.com/Fix-a-Squeaking-Bed-Frame	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos
How to Create Bus_pt	Português	http://www.wikihow.com/Create-Business-Letter-White-Space	.png	Página Web	procedimental/explicativo	fotografias, legendas, subtítulos
Como se tornar um acadêmico_pt	Português	http://www.wikihow.com/Become-an-Academic	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos
4 Ways to Clarify Butter_pt	Português	http://www.wikihow.com/Clarify-Butter	.png	Página Web	procedimental/explicativo	fotografias, legendas, subtítulos
25 anos após a independência_ng_pt	Português	http://www.nationalgeographic.com/photography/proof/2016/09/taji	.pdf	Página Web	informativa/picture story	fotografias, legendas, título, su
Lion_ng_pt	Português	http://kids.nationalgeographic.com/animals/lion/	.pdf	Página Web	informativa/educacional	fotografias, desenhos, título, si
Brazil dam burst engul_pt	Português	http://www.bbc.com/news/world-latin-america-34742272	.pdf	Página Web	informativa	fotografias, vídeos, legendas, t
Brexit_David Cameron to quit_pt	Português	http://www.bbc.com/news/uk-politics-36615028	.pdf	Página Web	informativa	fotografias, vídeos, mapa, lege
Islâmico apreende	Português	http://www.bbc.com/news/world-middle-east-32820857	.pdf	Página Web	informativa	fotografias, vídeos, mapa, lege
Sinai plane crash_Russia sus_pt	Português	http://www.bbc.com/news/world-middle-east-34687990	.pdf	Página Web	informativa	fotografias, mapa, legendas, tit

Fonte: Próprio autor.

A tabela⁵³ acima tem a função de facilitar a recuperação de determinadas informações de cada arquivo, além de proporcionar uma visão geral do corpus. Assim, a primeira seção trata de abordar o perfil geral do corpus analisado, e depois segue a análise por tipo de documento multimodal.

5.1 Análise geral

Um panorama geral da amostra analisada aponta para uma maioria de artigos retirados de páginas *web*, seguida de infográficos, manuais, e artigos acadêmicos. A

⁵² A base de dados do estudo está disponível para *download* por meio do link <https://www.dropbox.com/s/g7ky2traywqxbek/corpus.zip?dl=0>.

⁵³ A tabela completa está disponível em Apêndices.

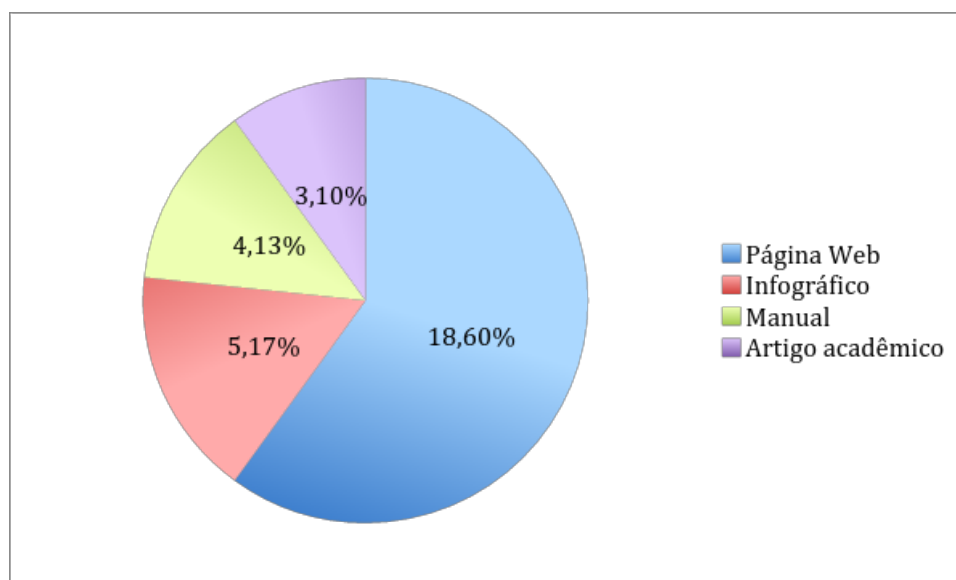
composição do corpus de análise utiliza como base alguns artigos do *website* *Wikihow* em língua inglesa, escolhidos de forma aleatória, e suas traduções geradas pelo Google Tradutor. A escolha por esses artigos parte dos achados do estudo piloto que apontam para um maior índice de incompatibilidade intersemiótica nesses artigos, especificamente por demonstrar relações entre parte das legendas de fotos ou imagens que constroem a narrativa de procedimento ou explicação de um processo. Portanto, o discurso de instrução/explicação multimodal (ou seja a instrução que utiliza texto e imagem em conjunto para se criar uma unidade semântica) também é conferida em menor escala em alguns manuais de produtos, e explicações (no sentido informacional) em outros *websites* de notícias, de entretenimento e educativos. Outros gêneros textuais tais como infográficos e artigos acadêmicos também são utilizados em menor escala para verificar se as incompatibilidades intersemióticas encontradas nessas páginas *web* também ocorrem nesses últimos.

É importante lembrar que este estudo não leva em consideração um balanceamento de número e tipo de documentos para se obter uma representação da amostra de análise. O intuito dessa amostra é expandir o número de documentos em que se encontrou incompatibilidades e adicionar outros tipos de documentos para se buscar e entender novos tipos de incompatibilidades intersemióticas.

Ainda que o desenho metodológico dessa investigação seja qualitativo, é

necessário demonstrar alguns números apenas para obter uma visão geral sobre a amostra analisada. Assim, a porcentagem dessa amostra pode ser visualizada a partir do gráfico a seguir.

Gráfico 1 - Perfil dos tipos de documentos multimodais



Fonte: Próprio autor.

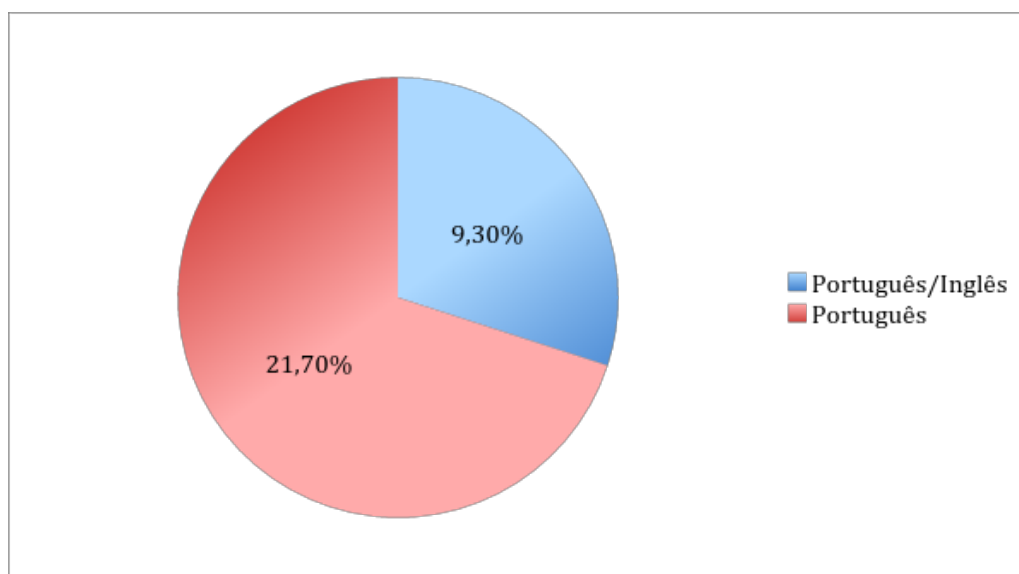
O gráfico acima apresenta uma amostragem total de 30 documentos multimodais com suas respectivas traduções. As cores representam os tipos de documentos multimodais (identificadas na legenda ao lado direito do gráfico), a numeração representa a quantidade de cada tipo de documento, seguida da porcentagem daquela determinada quantidade dentro do universo total da amostra.

A maior parte dos documentos é formada por artigos de páginas *web* (*Wikipedia*, *BBC*, *Buzzfeed*, e *National Geographic*) com 18 textos e 60% do total de documentos,

onde foram encontrados mais tipos de incompatibilidades intersemióticas. A seguir, os infográficos, também em formato digital e em sua totalidade disponibilizados em *websites*, computam 5 exemplares e um total de 17% da soma de todos os documentos. Em número aproximado, os manuais formam 4 unidades e 13% do total de documentos. Por fim, os dois artigos acadêmicos somam 10% do total de documentos multimodais.

O gráfico 2 a seguir apresenta a quantidade e porcentagem de documentos que foram analisados com o texto em português totalmente gerado pela tradução automática do Google Tradutor encontrado no navegador *Google Chrome* e pelos documentos com formato *pdf* ou imagem que necessitaram de uma intervenção manual⁵⁴ para serem analisados (portanto, textos originais no inglês com alguns trechos em português).

⁵⁴ Trechos de documento em formato de imagem onde houve suspeita de incompatibilidade intersemiótica foram transcritos e colados no Google Tradutor, e em seguida sua tradução para o português foi copiada e colada de volta no texto para registro de análise.

Gráfico 2 - Perfil de documentos traduzidos total e parcialmente

Fonte: Próprio autor.

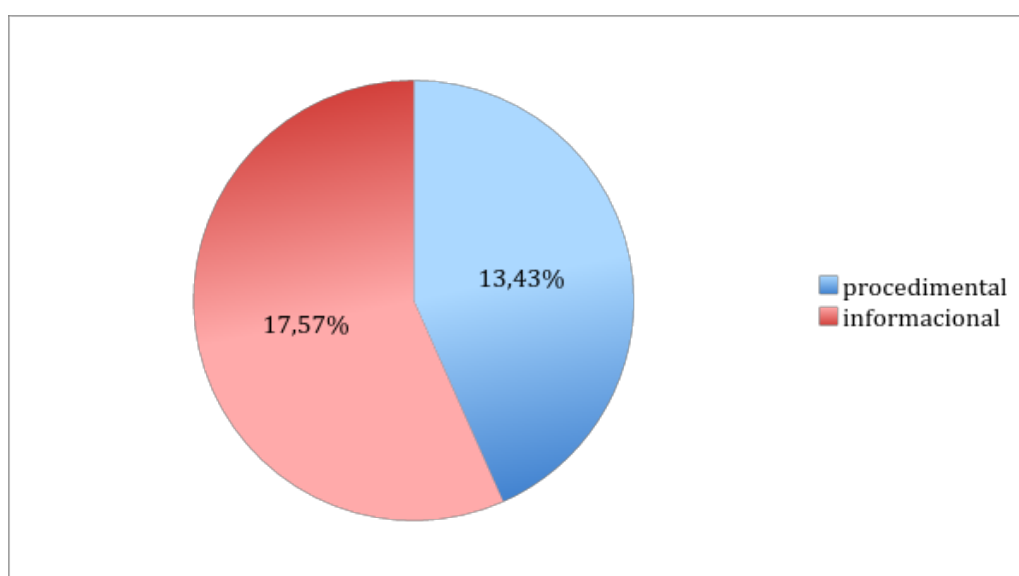
A maior parte de documentos inteiramente traduzidos pelo Google Tradutor compõe um total de 21 documentos correspondentes a 70% da amostra analisada. Já os documentos em inglês disponibilizados originalmente em pdf e em imagem (.png ou .jpg) e que mais tarde receberam uma tradução automática pontual correspondem a 30%, sendo 9 documentos analisados.

Este perfil é significativo porque há uma intervenção adicional ao processo de anotação das incompatibilidades intersemióticas em tais documentos. Conseqüentemente, há maior consumo de tempo de análise nesses documentos. Ao mesmo tempo em que apresentam maior potencial para variação semiótica, já que não

seguem uma estrutura rígida como um *template* nos artigos de página *web*. Contudo, esses casos serão tratados com maiores detalhes mais adiante.

O gráfico a seguir apresenta alguns dados sobre os tipos de discurso encontrados nos documentos analisados.

Gráfico 3 - Perfil de tipos de conteúdo dos documentos analisados



Fonte: Próprio autor.

Nota-se no gráfico 3 um maior número de documentos com conteúdo informativo, isto é, textos criados com a finalidade de informar o público acerca de um assunto específico, seja ele político, educacional, científico, ou recreativo. Tais documentos compreendem 57% do total, um pouco mais que os 43% de documentos com conteúdo procedimental. Esses foram em grande parte os artigos de instrução tais como os artigos do website *Wikihow* e os manuais.

Por fim, os três gráficos acima dão uma visão geral sobre os tipos de textos utilizados para a análise exploratória, e permitem generalizar sobre os fenômenos de incompatibilidade intersemiótica.

A subseção a seguir detalha a análise das páginas *web*.

5.2 *Webpages*

De acordo com Djonov (2007, p. 145, tradução nossa), “websites consistem em páginas *web* [*webpages*, em inglês], isto é, nódulos que incorporam elementos visuais, verbais e gradualmente sonoros e cinéticos, além de hiperlinks”. É notório que a página *web* se enquadra como um típico documento multimodal, e no caso das páginas analisadas não foi diferente. Contudo, apenas parte desse “nódulo” foi levada em consideração, qual seja a do escopo imagem-texto estabelecido no objetivo deste estudo.

Ainda de acordo com a autora (DJONOV, 2007, p. 145, tradução nossa) as funções das páginas *web* são de “apresentar conteúdo do *website*”, tipicamente na principal área de visualização, de “exibir a organização do conteúdo, das opções de navegação, e da posição do usuário dentro do *website*” (as questões relacionadas a usabilidade não foram consideradas neste estudo).

Dentro desses aspectos delimitados acima, pode-se encontrar algumas incompatibilidades intersemióticas na relação texto-imagem na área principal de visualização dos artigos disponíveis nas páginas *web*. Publicidades, caminhos de *hyperlink*, menus, *layout*, navegação e orientação não foram considerados.

A observação se atém em um primeiro momento na busca por possíveis erros de tradução automática no nível lexical e que tenham ao mesmo tempo algum tipo de relação semântica com alguma imagem na construção daquele determinado documento. Assim, o nível de análise é discursivo.

É crucial lembrar que neste primeiro momento a base utilizada como referência para os tipos de erros do par linguístico inglês-português é o TraVA (. (OLIVEIRA JR. et al, 2000; KAMEYAMA et. al., 1991). A seguir busca-se a relação intersemiótica de Liu e O'Halloran (2009), para verificar possíveis fenômenos intersemióticos.

A seguir especifica-se a análise das páginas da *Wikihow*.

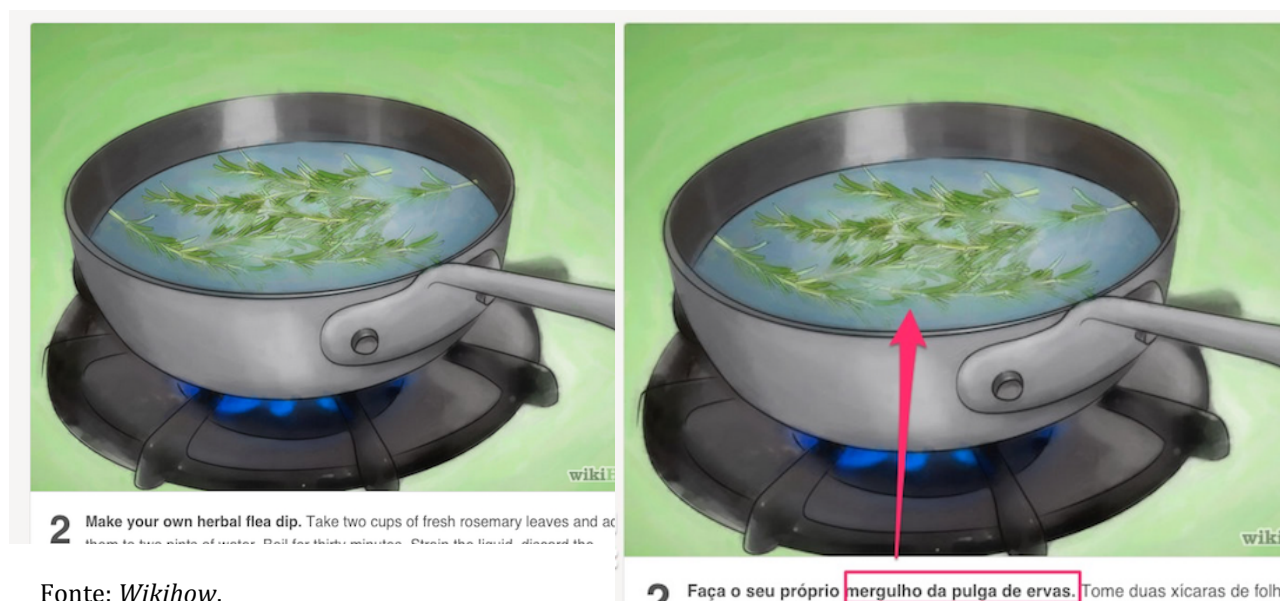
- ***Wikihow***

Ao todo, foram analisados 8 pares de artigos do *website* de instruções ilustradas. Os artigos apresentam em sua maioria uma estrutura fixa com unidades básicas na forma de fotografias, desenhos, legendas, subtítulos, e títulos. Do total de artigos analisados, metade apresentou uma média de cinco incompatibilidades intersemióticas, e a outra metade entre zero e uma (os dados podem ser examinados na tabela de

classificações nos apêndices da presente tese). Dentre o número total de incompatibilidades foi possível chegar a 4 tipos de categorias.

Além disso, observou-se que há uma repetição parcial das categorias encontradas no estudo piloto, sobretudo a incompatibilidade de ambiguidade intersemiótica. Em alguns casos há variações (e especificidades) do mesmo caso. Neles observa-se que há diferentes tipos de ambiguidade intersemiótica (geração de diferentes sentidos a partir da relação texto-imagem do erro de tradução automática) baseadas também nas diversas categorias de erro de tradução automática. A figura 12 abaixo ilustra a identificação da categoria de ambiguidade intersemiótica conforme encontrada no estudo piloto.

Figura 12 - Exemplo 1 de ambiguidade intersemiótica



Fonte: *Wikihow*.

As imagens da figura acima foram capturadas da versão original em inglês retirado de um artigo da *WikiHow* intitulado “How to get rid of fleas naturally”, em português uma possível tradução seria “Como se livrar de pulgas naturalmente”. À direita, está sua tradução para o português gerada pelo Google Tradutor. Em português a palavra “dip” pode ser traduzida como “solução”. Porém, o problema reside em utilizar a palavra “flea” (“pulga”, em português) de uma forma adjetivada, onde, em português essa construção é frequentemente utilizada com a palavra “contra”, resultando em “solução herbal contra pulgas”.

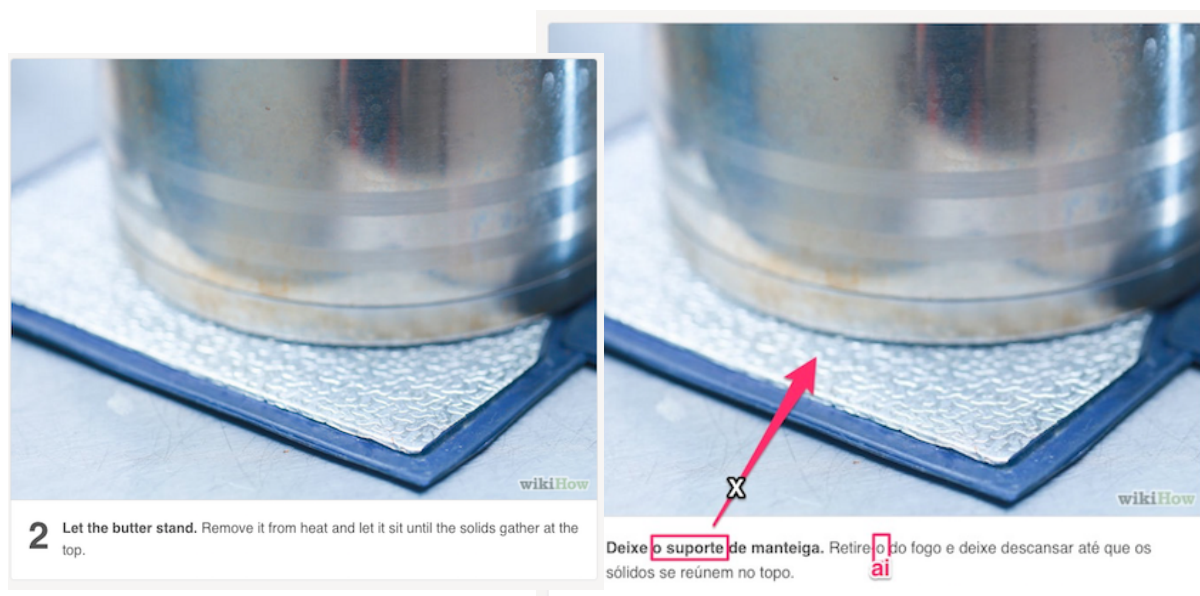
Os erros de tradução gerada pelo Google Tradutor em sua maioria partem de origem léxico-semântica, e portanto de sentido de "palavra incorreta", ou de "palavra desconhecida" (VILAR et al., 2006). Assim, a tradução gerada pela máquina, “mergulho da pulga de ervas”, ao descrever parte da ação representada pela figura acima pode criar uma "incompatibilidade de correspondência intersemiótica" (JONES, 2006).

A correspondência intersemiótica, diferentemente de um sinônimo ou repetição, caracteriza a relação entre um elemento textual e verbal com vistas para o uso conjugado entre significados verbais e visuais visando uma correspondência e a expansão do significado resultante (JONES, 2006, p. 194).

Outro tipo de ambiguidade intersemiótica é aquela encontrada a partir da escolha (automática) de outra palavra do mesmo grupo semântico mas que por um erro

de TA, coincidentemente, o sentido da palavra traduzida incorretamente se relaciona com outra parte da imagem, diferentemente do original. Baseado na categoria de "contingência intersemiótica" de Liu e O'Halloran, nomeia-se esse tipo de fenômeno de "incompatibilidade de contingência intersemiótica" conforme pode-se verificar na figura 13 abaixo:

Figura 13 - Exemplo 2 de ambiguidade intersemiótica



Fonte: *Wikihow*.

O exemplo da figura acima foi retirado do artigo "How to clarify butter" disponível no *website Wikihow*. No passo 2 podemos perceber que a descrição inicial em negrito utiliza o verbo "stand" no sentido de "repousar" normalmente visto em receitas, resultando numa possível tradução como "deixe a manteiga repousar". Contudo a

identificação de “stand” como substantivo gerou a palavra “suporte” em português, gerando uma nova conexão com a imagem, inexistente no original, pois na imagem há um suporte para a vasilha onde se encontra a manteiga derretida.

Esse fenômeno, chamado de incompatibilidade de contingência intersemiótica, aprofunda o mesmo caso de ambiguidade visto no artigo do *website BuzzFeed* descrito no estudo piloto, indicando um possível padrão.

Outro caso que se repete nos artigos da *WikiHow* são as incompatibilidades de relação temporal, como demonstra a figura a seguir:

Figura 14 - Exemplo 1 de incompatibilidade de relação intersemiótica de tempo





wikiHow

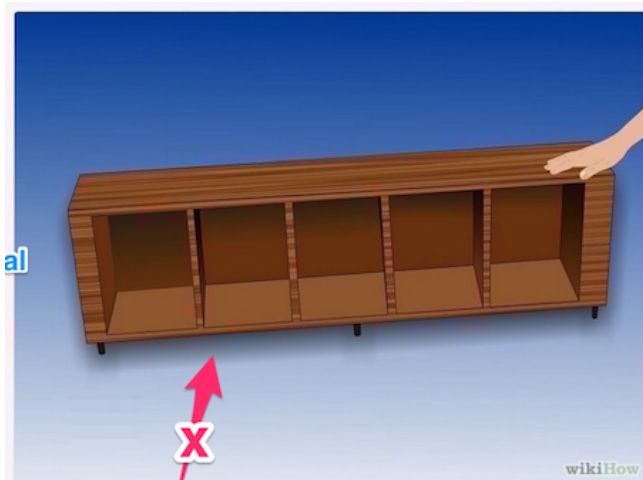
- 4** Turn it right side up. Turn the shelf so that it sits on its new feet. You now have a bench!



wikiHow

- 3** Adicionar rodas ou pernas. Vá para a sua loja de ferragem local e escolher algumas pernas para o seu banco. Você pode obter rodas (como o tipo que você vê em carrinhos de compras) ou pés dos móveis de madeira ou metal. O que quer que você sente combina com sua decoração e suas necessidades. Prenda os pés nos cantos do banco, conforme descrito nas instruções para o produto que você compra.

- Pelo menos quatro pés são necessárias. Seis são recomendados.
- Certifique-se o parafuso não interfira com as outras peças da prateleira segurando juntos. Seja estratégico!



wikiHow

- 4** Transformá-lo de cabeça para cima. Vire a prateleira de modo que ele se senta em seu novo pés. Agora você tem um banco!

Fonte: *Wikihow*.

As imagens da Figura 14 foram⁵⁵ capturadas do artigo “Ways to build a bench” no *Wikihow* em inglês, à esquerda; à direita, estão as imagens das mesmas partes do artigo, porém traduzidas para o português pelo Google Tradutor. A incompatibilidade de relação temporal se dá pela “quebra” da sequência narrativa, a partir de uma divergência léxico-semântica proporcionada pela palavra “transformar”. O passo três é importante para visualizar que o banco está prestes a ser “virado”, pois é possível ver a instalação dos pés do lado que será virado para baixo no passo 4. Dessa forma, o uso de “transformá-lo” ao invés de “virá-lo” (por exemplo) elimina a ação implicada no passo anterior.

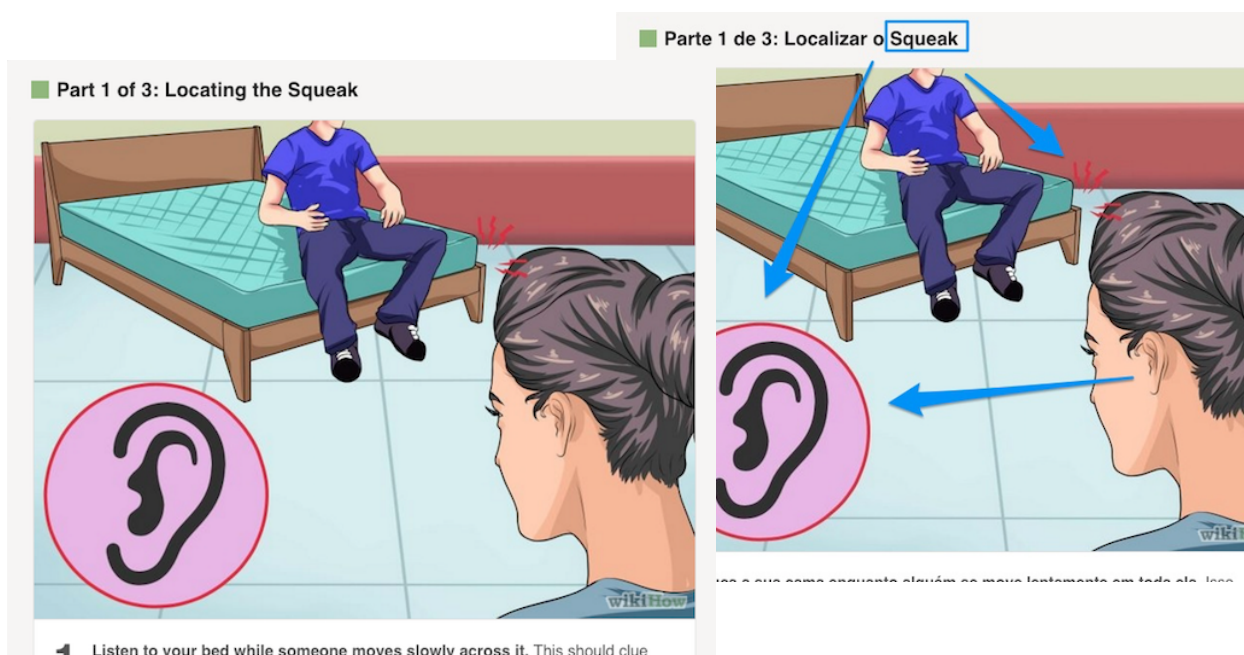
Conforme podemos observar, a incompatibilidade de relação temporal intersemiótica se dá quando há erro de TA e, coincidentemente, o sentido da palavra traduzida incorretamente se reconfigura. A partir daí, a relação temporal da imagem ou parte dela também se modifica semanticamente em sua relação com um evento representado por imagem ou texto naquele documento.

Além dos casos de ambiguidade e relação temporal, outro achado relevante nas incompatibilidades de relações texto-imagem são as formas verbais não identificadas pelo Google Tradutor. Isto é, de palavras que não foram traduzidas, e portanto mantidas

⁵⁵ Para melhor visualização acesse o a base de dados por meio do link <https://www.dropbox.com/s/g7ky2traywqxbek/corpus.zip?dl=0>.

em inglês. Nesse caso, a relação semântica da palavra com parte da imagem não se realiza. Este achado é substancialmente recorrente em muitos documentos analisados, como pode ser exemplificado na figura abaixo.

Figura 15 - Exemplo de incompatibilidade intersemiótica com palavras desconhecidas



Fonte: *WikiHow*.

Acima, encontra-se uma relação de incompatibilidade texto-imagem no artigo da *WikiHow* intitulado “How to fix a squeaking bed frame”. A anotação encontra-se em azul para identificar as incompatibilidades intersemióticas produzidas a partir de “palavras desconhecidas” (VILAR et. al., 2006) pelo Google Tradutor (ou que foram percebidas como tal, a partir do resultado visível da tradução). Assim, as setas azuis da anotação demonstram as relações entre a palavra “squeak” e os diferentes elementos visuais que

sugerem a tradução em português como “rangido”, e a localização do mesmo elemento semântico em imagem.

Conforme a classificação de erros de tradução proposta por VILAR et al. (2006, p. 698, tradução nossa), a categoria das palavras desconhecidas (*unknown words*) “podem ser distinguidas entre palavras verdadeiramente desconhecidas (ou o radical delas) e formas desconhecidas de radicais conhecidos”.

Estes foram os principais achados nos artigos da *Wikihow*, e que encontram-se presentes nos outros subgêneros de *websites*.

- **BBC**

Analisou-se um total de seis pares de artigos do portal de notícias da BBC. Em sua maioria os artigos abrangem o noticiário político e contém diversas mídias para composição do texto, tais como vídeos, infográficos, mapas e fotografias. Contudo, encontrou-se proporcionalmente menor número de incompatibilidades intersemióticas do que no *website Wikihow*: apenas dois casos de ambiguidade intersemiótica, conforme o exemplo ilustrado abaixo.

Figura 16 - Incompatibilidade intersemiótica em artigo da BBC



Fonte: BBC news.

Na figura acima podemos identificar novamente a categoria de "palavra desconhecida" (VILAR et. al., 2006) na tradução da sigla "PM" para *prime minister* (primeiro ministro). Mesmo que a sigla "PM" se mantenha em português, coincidentemente como "primeiro ministro", não há uso recorrente dessa forma na

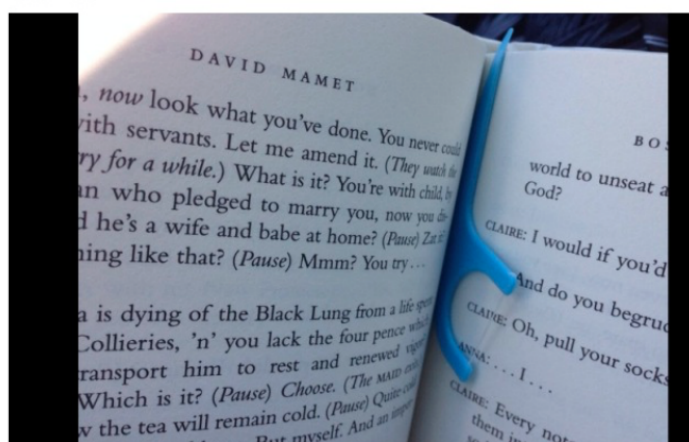
língua de chegada. Portanto a relação texto-imagem neste caso deixa de ter uma relação semântica explícita como no original.

- **Buzzfeed e National Geographic**

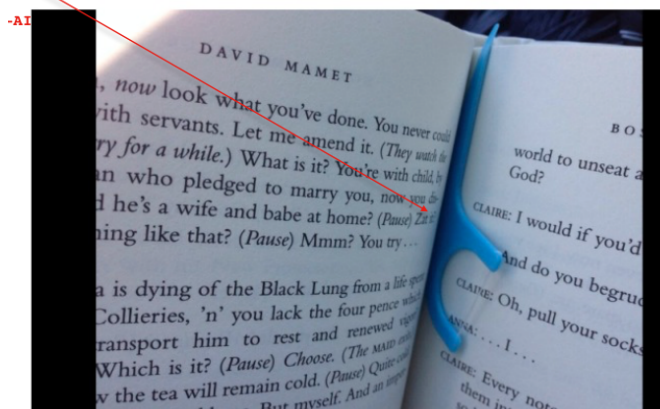
Por fim, os *websites* *Buzzfeed* e *National Geographic* apresentam poucas, porém significativas amostras de ambiguidades intersemióticas. No caso do sítio de entretenimento *Buzzfeed*, analisou-se 2 páginas *web*, das quais apenas uma apresentou incompatibilidades (total de 4, e todas do tipo ambiguidade intersemiótica). A narrativa do artigo intitulado “19 secrets Librarians will never tell you”, se utiliza de diversas mídias tais como desenhos, *gifs* animados, e fotografias para construir a unidade semântica em conjunto com os subtítulos e legendas de tom humorado. Dentre os achados, destaca-se o caso da ambiguidade intersemiótica gerada a partir de um erro de ordem de palavra na TA, conforme exemplo abaixo.

Figura 17 - Incompatibilidade intersemiótica em tradução automática

3. It's not uncommon to find ~surprises~ inside returned books — food, mold, and a few soaking wet books.



3. Não é incomum encontrar ~ surpresas ~ livros dentro devolvido - alimentos, mofo e alguns livros imersão wet.



Fonte: *Buzzfeed*.

A partir da figura acima, observa-se uma ambiguidade intersemiótica gerada a partir da categoria de erros de "ordem de palavras" (VILAR et. al., 2006). De acordo com a análise, tal problema é identificado a partir da falta de reconhecimento automático do símbolo “~” no lugar de aspas. Assim, escapa à TA a relação da imagem de um objeto (aparentemente um fio dental) entre as folhas de um livro com a nova ordem do bloco de palavras de que “dentro” faz parte.

Já no caso das duas páginas *web* do *National Geographic* houve um maior número de ambiguidade intersemiótica, sobretudo na página analisada da *National Geographic Kids*. O artigo trata de informações gerais sobre leões, de forma interativa e didática. É possível identificar diversas mídias como mapa, infográfico, desenhos e fotografias. Nesse documento pode-se notar novamente a categoria de contingência intersemiótica na relação demonstrada abaixo.

Figura 18 - Exemplo de incompatibilidade intersemiótica em página da *National Geographic Kids*



Fonte: *National Geographic Kids*.

Contudo, no caso acima há erro na tradução da palavra “lighter” por “claro” ao invés de “mais leve” (suprindo não só a definição adequada, mas a parte do sentido comparativo do adjetivo em questão) e, coincidentemente, o sentido da palavra traduzida incorretamente não se relaciona semanticamente com nenhuma parte da imagem, conforme demonstrado pelas setas vermelhas na anotação.

Assim, encerra-se a análise das páginas *web*. A próxima subseção trata da análise dos infográficos.

5.3 Infográficos

O infográfico intitulado *EU budget explained: expenditure and contribution by member state* da seção de notícias disponível no *website* do parlamento europeu ilustra quanto e para qual fim cada estado-membro da União Européia gasta seus recursos (no caso específico deste documento, a Alemanha foi selecionada). Neste infográfico não foi encontrado nenhum caso de incompatibilidade intersemiótica. Uma das possíveis

sugestões para tal achado é de que não há relação direta entre o significado do texto traduzido e parte da imagem (como o caso de uma legenda de um gráfico, por exemplo).

O infográfico informando a estrutura localizada na expansão do museu de Auckland⁵⁶ também segue o mesmo padrão do infográfico do sítio do parlamento europeu. Não há nenhuma incompatibilidade intersemiótica observada, pois não houve nenhuma divergência léxico-semântica que houvesse relação também em nível semântico com alguma parte da representação da estrutura do museu.

O mesmo caso ocorre nos infográficos *More than honey: honeybees and our food system*, que é parte de um artigo do portal de notícias do jornal *The Huffington Post*, e *Cool truths about the tooth* da *Televox*. Os espaços onde texto e imagem estão mais envolvidos não apresentam qualquer incompatibilidade intersemiótica.

Já o caso do infográfico intitulado *A breakdown of a person's google results*, a campanha da empresa de marketing chama a atenção do leitor para o “ranqueamento” de perfis pessoais na primeira página dos resultados do motor de busca do Google. E já no subtítulo podemos perceber uma incompatibilidade de ambiguidade intersemiótica, conforme ilustra a figura 19 abaixo.

⁵⁶ Vide base de dados do estudo.

Figura 19 - Infográfico da empresa *Brandyourself* com incompatibilidade intersemiótica



Fonte: *Brandyourself*.

Conforme demonstra a figura acima, as palavras *look* encontradas no subtítulo referem-se a como as pessoas se parecem quando são encontradas na página de resultados do motor de busca do Google. Contudo, não foi esse o sentido encontrado na tradução automática no Google Tradutor. Como pode-se notar, na tradução automática exibida em letras vermelhas⁵⁷ o verbo gerado foi “olhar”, o que proporcionou uma incompatibilidade de “relação intersemiótica de contingência”, pois formou-se, por

⁵⁷ As letras vermelhas representam trecho do texto que foi transcrito para a TA do Google e colado de volta no documento para análise.

coincidência uma relação do novo verbo com a ação do sujeito na imagem olhando para a tela de um computador.

A seguir, analisa-se as incompatibilidades encontradas nos manuais.

5.4 Manuais

O manual da *Airbus* intitulado *Aircraft characteristics – airport and maintenance planning* demonstra com detalhes as características do avião para manutenção. Contudo, este manual apresenta apenas a terminologia ligada às partes do avião, e por esse motivo as relações de imagem e texto não foram analisadas. Buscou-se outras relações onde houvesse a construção frasal, ou que a terminologia fosse parte da oração, mas tais casos não foram encontrados.

O mesmo caso acontece com o manual da máquina fotográfica *Canon*, modelo *EOS 1100D Rebel T3*. Contudo, neste exemplo específico há uma série de casos de instrução, com sequências criadas com texto e imagem, ao invés de utilizar apenas terminologia para identificar componentes do produto.

No manual do usuário da câmera *Hero 3 – Black Edition* também não observou-se incompatibilidades intersemióticas. Porém, notou-se um expressivo uso de ícones e símbolos que representam funcionalidades da câmera como parte integrante dos textos que davam instruções. Mas nos exemplos onde encontrou-se relações intersemióticas

entre imagem representando o produto e texto referindo-se a ela não houve incompatibilidades dentro do escopo de investigação.

O manual do usuário do tablete da *Samsung Galaxy TabS* tampouco apresenta incompatibilidades intersemióticas. As relações texto-imagem identificadas na estrutura de organização informacional apresentada neste documento exibe instruções em forma de frases compostas por ícones do aparelho. Portanto, não se enquadra no escopo delimitado nesta investigação.

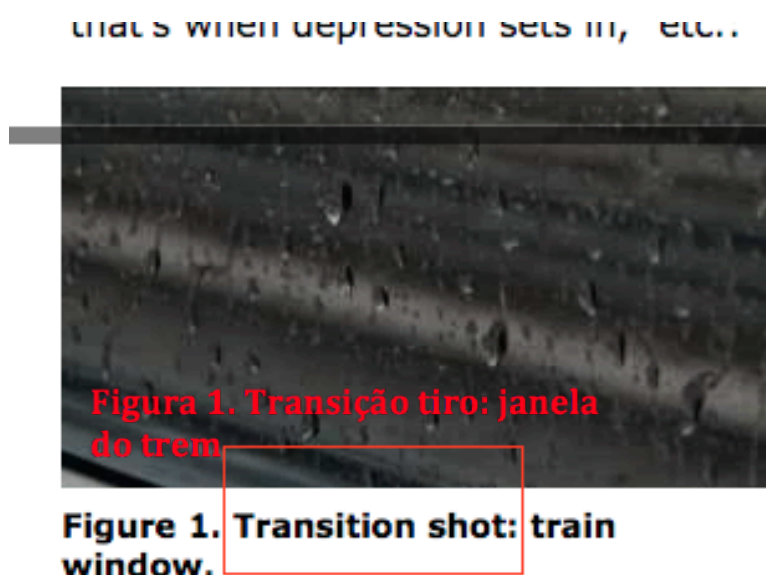
5.5 Artigos científicos

Ao todo foram analisados três artigos científicos. Dois deles não apresentam incompatibilidade intersemiótica, quais sejam *Rhetorical relations in multimodal documents* de Taboada e Habel (2013) e *A framework for the multimodal analysis of online news galleries: What makes a “good” picture gallery?* de Caple e Knox (2015). Há uma série de relações intersemióticas especialmente entre legenda e figura, mas não houve nenhuma incompatibilidade intersemiótica causada, a princípio, por erro de tradução automática.

Porém, o artigo de Romero-Fresco (2013) intitulado *Accessible filmmaking: Joining the dots between audiovisual translation, accessibility and filmmaking* exibe uma

incompatibilidade intersemiótica entre uma figura e sua legenda, conforme podemos identificar na figura 20 a seguir.

Figura 20 - Incompatibilidade intersemiótica em artigo acadêmico



Fonte: Romero-Fresco (2013).

A figura acima exibe uma captura de imagem onde se vê parte da "figura 1" do artigo e sua legenda abaixo *Transition shot: train window*, ou seja "plano de transição: janela do trem" em uma possível tradução para o português. A figura acima foi capturada de documento em *pdf*, que foi analisado a partir de transcrições de determinados trechos diretamente no Google Tradutor. Contudo, a tradução automática do Google gerou "tiro" ao invés de "plano", contribuindo para uma incompatibilidade intersemiótica de ambiguidade com a imagem da Figura 20.

A partir da ideia de multiplicação de sentido (BATEMAN, 2014), os achados da presente pesquisa contribuem para demonstrar que os erros de tradução automática podem reconfigurar total ou parcialmente, ou seja, além de multiplicar, podem “subtrair” e “dividir” a relação intersemiótica entre os significados entre texto e imagem em documentos multimodais. Um exemplo disso parte das categorias de palavras desconhecidas, ou seja, que permanecem em inglês na tradução.

Por fim, este capítulo tratou da análise geral da base de dados, assim como cada tipo de documento, exemplificando os casos mais representativos. A discussão dos achados, assim como reflexões e pesquisas futuras são tratadas no capítulo subsequente.

6. Considerações finais

Por meio do subsídio teórico da "multimodalidade", mais especificamente do modelo de "textura intersemiótica" proposto por Liu e O'Halloran (2009), e da "linguística computacional", mais especificamente da "tipologia de erros de tradução automática" baseada em Vilar et. al. (2006) e do conceito de "incompatibilidade de tradução" (*translation mismatch*, em inglês) de Kameyama et. al (1991), o presente estudo intitulado "ampliando olhares sobre a tradução automática online: um estudo exploratório de categorias de erros de máquina de tradução gerados em documentos multimodais" identificou e categorizou relações semânticas entre texto e imagem em documentos multimodais traduzidos automaticamente para a avaliação de tradução automática.

A partir dos achados produzidos durante o estudo piloto foi possível identificar e caracterizar o fenômeno da "incompatibilidade intersemiótica", ou seja, divergências semânticas entre os componentes linguísticos e visuais geradas a partir de erros de TA.

No estudo piloto, tais incompatibilidades, respectivamente "ambiguidade intersemiótica" e "relações temporais intersemióticas" foram encontradas em páginas *web* de instruções e de portal de notícias. Para a composição da amostra principal de análise optou-se por explorar e catalogar maior variedade de tipos de categorias. Para

tanto, decidiu-se aumentar o número de páginas *web* de instruções e utilizar manuais de produtos para investigar se o gênero textual de instrução contém os mesmos tipos de incompatibilidades intersemióticas. Identificou-se o mesmo processo a partir do aumento de artigos de *websites* de notícias e artigos acadêmicos. A inclusão de infográficos na base de dados analisada e se justifica pela sua natureza de relação texto entre imagem mais “fluida”, no sentido de multiplicar a informação pela relação mais intrínseca entre os dois componentes de sentido (BATEMAN, 2014).

De um total de 30 pares de textos, 18 foram páginas *web* dos *websites* *Wikihow*, *BBC*, *Buzzfeed*, e *National Geographic Kids*, 5 foram infográficos de diversos gêneros, 4 foram manuais de produtos eletrônicos, e 3 pares de textos foram artigos acadêmicos. Todos os documentos originalmente no inglês foram traduzidos automaticamente para o português, por meio do serviço online Google Tradutor.

A análise apontou para uma maior proporção e tipos de "incompatibilidades intersemióticas" nas páginas *web*. Houve um menor número representado nos infográficos e nos artigos acadêmicos. Não foi encontrada nenhuma relação intersemiótica nos manuais, o que se encaixa no escopo do presente trabalho (por exemplo, a maior parte das relações se dá dentro do escopo terminológico – em nomes de funções e botões - dentro do componente linguístico, fora de uma oração ou relação coesiva discursiva).

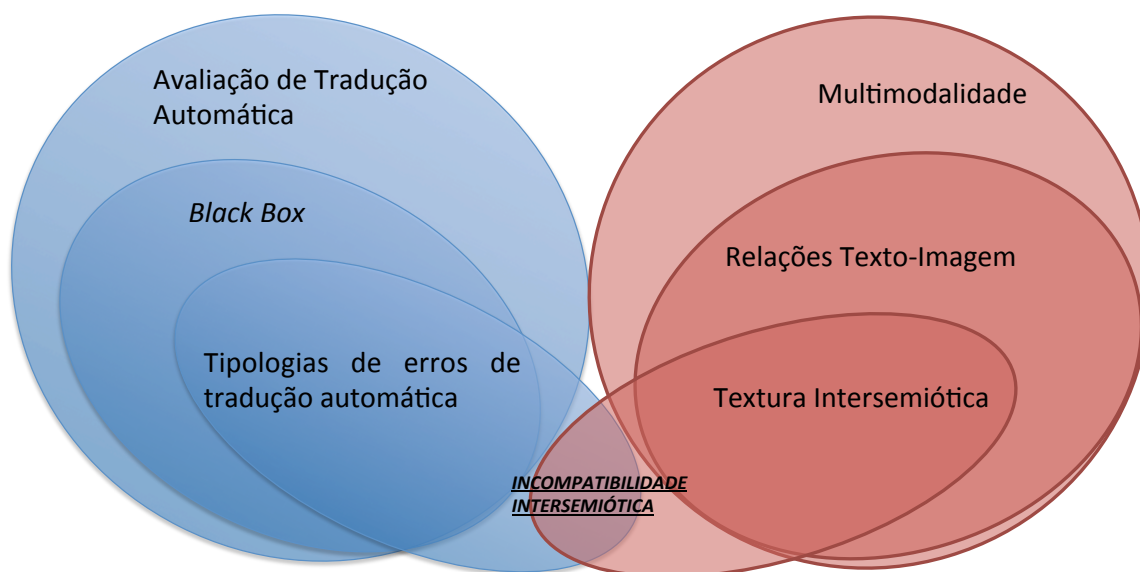
Verificou-se que os dois tipos de incompatibilidades intersemióticas encontradas no estudo piloto se repetem na análise, mas com maior número de especificidades como as chamadas “ambiguidades intersemióticas”, que foram identificadas em categorias mais específicas, como as de "contingência" e de "correspondência" (LIU; O'HALLORAN, 2009).

Também observou-se um padrão de problemas de escolha léxico-semântica, de “palavras desconhecidas” e de “ordem de palavras” dentro da tipologia de erros de tradução automática (RIBEIRO, 2006; VILAR et al., 2006) . Tais problemas foram relacionados com partes da imagem a qual se referem, mais especificamente dentro do modo de legenda descrevendo parte da imagem.

Dessa forma, o presente estudo responde a pergunta de pesquisa: **“Como se dão as relações texto-imagem em documentos multimodais a partir dos resultados da tradução automática?”** - em incompatibilidades intersemióticas de ambiguidade, contingência, e tempo a partir das categorias de formas incorretas e desambiguação incorreta. Tal achado corrobora e aprofunda a catalogação das incompatibilidades intersemióticas encontradas no estudo piloto. Os modelos de tipologia de erros de tradução e textura multimodal propostos foram satisfatórios para identificar as relações texto-imagem encontradas na análise.

Neste contexto, podemos identificar o fenômeno da incompatibilidade intersemiótica na interface teórico-analítica proposta no capítulo "Marco Teórico" abaixo:

Figura 21 - Proposta de incompatibilidade intersemiótica na interface de investigação



Fonte: Próprio autor.

O presente estudo também responde a seguinte pergunta de pesquisa: **“Como a abordagem da multimodalidade contribui para a avaliação de resultados de tradução automática em documentos multimodais?”** - Na identificação e catalogação de relações intersemióticas que escapam da função da máquina e da expansão do componente intersemiótico na tipologia de erros de tradução automática.

A utilização de ambas as perspectivas associadas dentro da ciência da informação também contribuem para os estudos utilizando multimodalidade e informação semântica (HJORLAND, 2007) dentro da área da Organização da Informação e do Conhecimento. Esta contribuição se realiza na medida em que oferece uma abordagem teórico-metodológico para a análise de relações entre símbolos e palavras dentro do contexto de aplicação de uma ferramenta semântica (HJORLAND, 2007), qual seja, de um tradutor automático.

A próxima subseção trata das limitações da pesquisa.

6.1 Limitações da pesquisa

Uma das limitações da pesquisa se dirige ao escopo linguístico empregado. A unidade semântica do componente textual limitou-se a buscar léxicos empregados em orações e suas ligações coesivas com parte do componente visual. Assim, o escopo de análise não se estendeu ao impacto que tais relações poderiam ter na narrativa como um todo (exceto nos casos de relações intersemióticas temporais, onde o alcance das sequências se estende para outras partes do texto). Desta forma, não se levou em consideração a relação das incompatibilidades intersemióticas no discurso e gênero textual (especialmente levando em consideração seus contextos de produção) utilizado.

A variedade de gêneros textuais também foi um fator limitador da pesquisa. Utilizou-se apenas algumas variedades de páginas web, infográficos, artigos acadêmicos e manuais de produtos eletrônicos.

O instrumental empregado para anotar as relações intersemióticas também foi limitador. Não encontrou-se durante a pesquisa tecnologia que disponibilizasse automática ou semi-automaticamente a anotação das relações texto-imagem, tampouco a quantificação delas. A falta de um programa que desempenhasse tais funções gerou um dispêndio substancial de tempo e energia para o exame e anotação manual das relações, especialmente nos documentos em que se impossibilitava o reconhecimento textual de forma automática pelo Google Tradutor.

Com base nessas limitações formulam-se as seguintes recomendações para pesquisas futuras.

6.2 Sugestões e recomendações

Com base nas limitações descritas acima, o presente estudo propõe as seguintes sugestões para pesquisas futuras:

- Expandir a pesquisa quantitativamente, isto é, aumentar o número de tipos de documentos para a análise de pares linguísticos, e identificar maior número de incompatibilidades intersemióticas em erros de tradução automática.

- Expandir a pesquisa qualitativamente: incluir dimensões discursivas e gêneros textuais que apresentem maior multiplicidade de sentido com base na relação texto-imagem como vídeos e histórias em quadrinhos, por exemplo. E analisar as configurações de incompatibilidades intersemióticas com base no contexto informacional de um determinado gênero textual.
- Formalizar e classificar as variedades de incompatibilidades intersemióticas por meio da elaboração de esquemas de anotação visando posteriores testes de avaliação de tradução automática.
- Desenvolver *softwares* capazes de disponibilizar a anotação automática e semi-automática de tais relações.
- Realizar testes de usabilidade e ergonomia com este (Google Tradutor) e outros sistemas de tradução automática a partir de incompatibilidades intersemióticas, visando a melhoria desses sistemas.
- Explorar possibilidades de aplicação desta interface teórico-metodológica em contextos de acessibilidade e geração de tradução automática.

7. Referências

ALMEIDA, C. C. DE. **Elementos de linguística e semiologia na organização da informação**. São Paulo: Cultura Acadêmica, 2011.

ALPAC. **Languages and machines: computers in translation and linguistics**. Washington, D.C.: A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, 1966.

BARANOW, U. G. Perspectivas na contribuição da linguística e de áreas afins a ciência da informação. **Ciência da Informação**, v. 12, n. 1, p. 23–35, 1983.

BATEMAN, J. A. **Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents**. New York: Palgrave MacMillan, 2008.

BATEMAN, J. A. **Text and Image: A Critical Introduction to the Visual/Verbal Divide**. New York: Routledge, 2014.

BELKIN, N. Information concepts for information science. **Journal of Documentation**, v. 34, p. 55–85, 1978.

BOITET, C. et al. **MT on and for the Web**. Proceedings of IEEE NLP-KE '10. **Anais...** In: IEEE NLP-KE '10. Beijing, China: 2010

BORKO, H. Information science: What is it? **American Documentation**, v. 19, n. 1, p. 3–5, jan. 1968.

BORODO, M. Multimodality, translation and comics. **Perspectives: Studies in Translatology**, v. 23, n. 1, p. 22–41, 2015.

BOUGNOUX, D. Sciences de l'information et de la communication. In: **Textes essentiels**. [s.l.] Larousse, 1993.

BRANCO, ET AL. **The Portuguese Language in the Digital Age**. [s.l.] Springer, 2012.

BROOKES, B. C. The foundations of Information Science: Part I. Philosophical aspects. **Journal of Information Science**, v. 2, n. 3–4, p. 125–133, 1980.

CALUDE, A. Machine Translation of Various Text Genres. **Te Reo; the New Zealand Linguistic Society Journal**, v. 46, p. 67–94, 2004.

CAPLE, H.; KNOX, J. S. A framework for the multimodal analysis of online news galleries: What makes a “good” picture gallery? **Social Semiotics**, v. 25, n. 3, p. 292–321, 2015.

CAPURRO, R.; HJORLAND, B. The concept of information. **Annual review of Information Science and Technology**, v. 37, p. 343–411, 2003.

CRESWELL, J. **Research Design Qualitative, Quantitative, and Mixed Methods Approaches**. 3rd. ed. California: Sage, 2009.

DAYRELL, C. **Sistemas de tradução automática: Avaliação e propostas de melhoria**. Dissertação de mestrado—Belo Horizonte: UFMG, 1999.

DJONOV, E. Website hierarchy and the interaction between content organization, webpage and navigation design: A systemic functional hypermedia discourse analysis perspective. **Information Design Journal**, v. 15, n. 2, p. 144–162, 2007.

DORR, B. **SOLVING THEMATIC DIVERGENCES IN MACHINE TRANSLATION**. Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics. **Anais...**Pittsburgh, Pennsylvania, USA: Association for Computational Linguistics, jun. 1990Disponível em: <<http://www.aclweb.org/anthology/P90-1017>>

DORR, B. Machine Translation Divergences: A Formal Description and Proposed Solution. **Computational Linguistics**, v. 20, n. 4, p. 597–634, 1994.

DUQUE, C. G. SiRILiCO uma proposta para um sistema de recuperação de Informação baseado em teorias da linguística computacional e ontologia. **Perspectivas em Ciência da Informação**, v. 10, p. 253–258, 2005.

FARRADANE, J. Towards a true information science. **Information Scientist**, v. 10, p. 91–101, 1976.

FARRADANE, J. The nature of information. **Journal of Information Science**, v. 1, n. 3, p. 13–17, 1979.

GALVEZ, C. Aplicación de transductores de estado-finito a los procesos de unificación de términos. **Ciência da Informação**, v. 35, n. 3, p. 67–74, 2006.

GONZALEZ DE GOMEZ, M. N. Novas fronteiras tecnológicas das ações de informação: questões e abordagens. **Ibict**, v. 33, n. 1, p. 55–67, 2004.

GONZALEZ DE GOMEZ, M. N. Luciano Floridi e os problemas filosóficos da informação. **IncID: Revista de Ciência da Informação e Documentação**, v. 4, p. 3–29, 2013.

HALLIDAY, M. A. K. **Language as social semiotic: towards a general sociolinguistic theory**. Columbia: Hornbeam Press, 1975.

HALLIDAY, M. A. K. **Halliday: System and Function in Language. Selected papers edited by Gunther Kress**, 1976.

HALLIDAY, M. A. K. **Language as social semiotic: the social interpretation of language and meaning**. London: Arnold, 1978.

HALLIDAY, M. A. K. **Introduction to Functional Grammar**. Londres: Arnold, 1985.

HALLIDAY, M. A. K. **Introduction to Functional Grammar**. 2. ed. Londres: Arnold, 1994.

HALLIDAY, M. A. K.; HASAN, R. **Cohesion in English**. London: Longman, 1976.

HALLIDAY, M. A. K.; MATTHIESSEN, C. **An introduction to functional grammar**. [s.l.] Arnold, 2004.

HIIPPALA, T. **Modelling the structure of a multimodal artefact**. Thesis—Helsinki: University of Helsinki, 2013.

HJORLAND, B. **Knowledge organization systems (KOS) Concepts**, 2007.

HJORLAND, B. What is Knowledge Organization (KO)? **Official Quarterly Journal of the International Society for Knowledge Organization**, v. 35, n. 2/3, p. 86–101, 2008.

HODGE, R.; KRESS, G. **Social Semiotics**. Cambridge: Polity, 1988.

HUTCHINS, W. J. **Machine Translation: Past, Present, Future**. New York, NY, USA: John Wiley & Sons, Inc., 1986.

HUTCHINS, W. J. **Recent developments in machine translation: a review of the last five years**. (T. W. Dan Maxwell Klaus Schubert, Ed.)New directions in machine translation. **Anais...**Foris Publications, ago. 1988

HUTCHINS, W. J. **Research methods and system designs in machine translation: a ten-year review, 1984-1994**. (D. Clarke, A. Vella, Eds.)Machine Translation: Ten Years On. **Anais...**Cranfield University Press, 1994

HUTCHINS, W. J. **First steps in mechanical translation**. (V. Teller, B. Sundheim, Eds.)MT Summit VI: past, present, future. **Anais...**1997

HUTCHINS, W. J. **The development and use of machine translation systems and computer-based translation tools.** (C. Zhaoxiong, Ed.) International Conference on Machine Translation & Computer Language Information Processing. **Anais...** Beijing, China: 1999

HUTCHINS, W. J. Early Years in Machine Translation: Memoirs and biographies of pioneers. In: Amsterdam: John Benjamins, 2000. p. 299–312.

HUTCHINS, W. J. Machine Translation over Fifty Years. **Histoire, Epistemologie, Langage, Tome XXII**, v. 23, p. 7–31, 2001.

HUTCHINS, W. J. **Towards a definition of Example-based Machine Translation.** Proceedings of the Example-based Machine Translation. **Anais...** Phuket, Thailand: 2005

HUTCHINS, W. J. Encyclopedia of Language and Linguistics. In: Second ed. Oxford: Elsevier, 2006. v. 7p. 375–383.

HUTCHINS, W. J. Machine translation: A concise history. **Journal of Translation Studies**, v. 13, n. 1–2, p. 29–70, 2010.

HUTCHINS, W. J. The Routledge Encyclopedia of Translation Technology. In: New York: Routledge, 2015. p. 120–136.

IEDEMA, R. Multimodality, resemiotization: Extending the analysis of discourse as multi-semiotic practice. **Visual Communication**, v. 2, n. 1, p. 29–57, 2003.

JEWITT, C. **The Routledge Handbook of Multimodal Analysis.** London: Routledge, 2009.

JONES, J. **Multiliteracies for academic purposes: A metafunctional exploration of intersemiosis and multimodality in university textbook and computer-based learning resources in science.** Tese de doutorado—Sidney, Austrália: University of Sidney, 2006.

KALTENBACHER, M. Perspectives on Multimodality: From the early beginnings to the state of the art. **Information Design Journal**, v. 12, n. 3, p. 190–207, 2004.

KAMEYAMA, M.; OCHITANI, R.; PETERS, S. **Resolving Translation Mismatches With Information Flow.** Proceedings of the Annual Meeting of the Association for Computational Linguistics ACL91. **Anais...**1991

KARPATSCHOFF, B. **Human Activity Contributions to the Anthropological Sciences From a Perspective of Activity Theory**. Copenhagen: Dansk psykologisk Forlag, 2000.

KOEHN, P. **Statistical Machine Translation**. UK: Cambridge, 2010.

KRESS, G. The Routledge handbook of multimodal analysis. In: London: Routledge, 2009.

KRESS, G.; VAN LEEUWEN, T. **Reading Images: The grammar of visual design**. 1st. ed. London: Routledge, 1996.

KRESS, G.; VAN LEEUWEN, T. **Multimodal Discourse: The modes and media of contemporary communication**. London: Edward Arnold, 2001.

KRESS, G.; VAN LEEUWEN, T. **Reading Images: The Grammar of Visual Design**. 2nd. ed. London: Routledge, 2006.

LAGARDA, A. L. et al. **Statistical Post-editing of a Rule-based Machine Translation System**. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. **Anais...: NAACL-Short '09**. Boulder, Colorado: Association for Computational Linguistics, 2009. Disponível em: <<http://dl.acm.org/citation.cfm?id=1620853.1620913>>

LEMKE, J. Multiplying Meaning: Visual and Verbal Semiotics in Scientific Text. In: **Reading Science**. Londres: Routledge, 1998. p. 87–113.

LIU, Y.; O'HALLORAN, K. L. Intersemiotic Texture: analyzing cohesive devices between language and images. **Social Semiotics**, v. 19, n. 4, p. 367–388, 2009.

LLITJÓOS, A. F.; CARBONELL, J. G.; LAVIE, A. **A framework for interactive and automatic refinement of transfer-based machine translation**. Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT). **Anais... In: ANNUAL CONFERENCE OF THE EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION (EAMT)**. Budapeste: 2005

MAIA, B.; BARREIRO, A. Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português. In: **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**. Lisboa, Portugal: IST press, 2007. p. 205–216.

MANN, W. C.; THOMPSON, S. A. Rhetorical Structure Theory: Toward a functional theory of text organization. **Text**, v. 8, n. 3, p. 243–281, 1988.

MARCUSCHI, L. A. Gêneros textuais: definição e funcionalidade. In: **Gêneros textuais e ensino**. Rio de Janeiro: Lucena, 2002. p. 20–36.

MARTIN, J. R. **English text: system and structure**. [s.l.] John Benjamins Pub. Co, 1992.

MARTINEC, R.; SALWAY, A. A system for image-text relations in new (and old) media. **Visual Communication**, v. 4, n. 3, p. 337–371, 2005.

MENDONÇA, E. S. A Linguística e a Ciência da Informação: estudos de uma intersecção. **Cincia da Informação**, v. 29, n. 3, p. 50–70, 2000.

MOLLICA, M. C.; GONÇALEZ, M. **Linguística e Ciência da Informação: diálogos possíveis**. Curitiba: APPRIS, 2012.

NUNES, M. G. V. **Processamento de línguas naturais: para quê e para quem**. [s.l.] Notas Didáticas ICMC-USP, 2008.

OCH, F. J. **Statistical Machine Translation: Foundations and Recent Advances**. The Tenth Machine Translation Summit. **Anais...** In: THE TENTH MACHINE TRANSLATION SUMMIT. Phuket, Thailand: 2005

O'HALLORAN, K. L. **Mathematical discourse: Language, symbolism and visual images**. London: Continuum, 2005.

OLIVEIRA JR., O. et al. **A crytical analysis of the performance of English-Portuguese-English MT systems**. Anais do V Encontro para o Processamento Computacional da língua portuguesa escrita e falada. **Anais...** In: PROPOR. Atibaia, SP: 2000

O'TOOLE, M. **The language of displayed art**. London: Leicester University Press, 1994.

PIRES, T. B.; DUQUE, C. G. Sistemas de gerenciamento de tradução: uma proposta de análise multimodal. **DataGramZero**, v. 16, n. 3, 2015.

RIBEIRO, G. C. B. **Avaliação de tradução automática no mercado de localização de software: um estudo de caso**. Dissertação de mestrado—Rio de Janeiro: PUC-Rio, 2006.

RICOY, R. DE P. Multimodality in translation: Steps towards socially useful research. **Multimodal Communication**, v. 1, n. 2, p. 181–203, 2012.

ROMERO-FRESCO, P. Accessible Filmmaking: Joining the dots between audiovisual translation, accessibility and filmmaking. **Journal of Specialized Translation**, v. 20, p. 201–223, 2013.

ROYCE, T. Synergy on the Page: Exploring intersemiotic complementarity in page-based multimodal text. **JASFL Occasional papers**, v. 1, n. 1, p. 25–49, 1998.

ROYCE, T. New Directions in the analysis of Multimodal Discourse. In: **New directions in the analysis of multimodal discourse**, New York: Routledge, 2007. p. 63–109.

SAMPAIO, C. DA S. **ARQUITETURA MULTIMODAL DA REPRESENTAÇÃO DA GEOINFORMAÇÃO: UMA PROPOSTA**. Tese de doutorado não publicada—Brasília: UnB, 2016.

SANTOS, D. Um Centro de Recursos para o Processamento Computacional do português. **DataGramZero**, v. 3, n. 1, 2002.

SANTOS, D. **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**. Lisboa, Portugal: IST Press, 2007.

SANTOS, D.; MAIA, B.; SARMENTO, L. **Gathering empirical data to evaluate MT from English to Portuguese**. Proceedings of LREC 2004. **Anais...** In: WORKSHOP ON THE AMAZING UTILITY OF PARALLEL AND COMPARABLE CORPORA. Lisboa, Portugal: 2004

SANTOS, M. A. DOS. **Interface multimodal de interação humano-computador em sistema de recuperação de informação baseado em voz e texto em português**. Dissertação de mestrado—Brasília: UNB, 2013.

SARMENTO, L. Ferramentas para experimentação, recolha e avaliação de exemplos de tradução automática. In: **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**. Lisboa, Portugal: IST press, 2007. p. 193–203.

SILVA, F. **Análise comparativa dos resultados de mecanismos de tradução automática baseados em regras e estatística**. Dissertação de mestrado—Florianópolis: UFSC, 2010.

SOUSA, E. E. DE. **A organização da informação e o ensino técnico do DF: Um modelo para promover o feedback para professores nas salas de aula**. Brasília: UnB, 2015.

STEINMETZ, E. H. R. **A contribuição da Arquitetura da Informação na construção e utilização de Ambientes Informacionais Colaborativos de Ensino/Aprendizagem.** Brasília: UnB, 2015.

TABOADA, M.; HABEL, C. Rhetorical relations in multimodal documents. **Discourse**, v. 15, n. 1, p. 65–89, maio 2013.

TRIPATHI, S.; SARKHEL, J. K. **Approaches to machine translation.** Annals of Library and Information Studies. **Anais...**2010

VIGOTSKY, L. S. **Mind in society: The development of higher psychological processes.** Cambridge, Mass.: Harvard University Press, 1978.

VILAR, D. et al. **Error analysis of statistical machine translation output.** . In: PROCEEDINGS OF LREC. Genoa: 2006

WHITE, J. S. How to evaluate machine translation. In: **Computers and Translation: A translator's guide.** Amsterdam/Philadelphia: John Benjamins Publishing, 2003. v. 35p. 211–244.

ANEXO A – Parecer do co-orientador estrangeiro



☒ Universität Bremen · Fachbereich 10 · Postfach 33 04 40 · 28334 Bremen

Coordenação de Aperfeiçoamento de Pessoal de
Nível Superior - CAPES
Brasil

Ihr Zeichen:

Ihre Nachricht vom:

Unser Zeichen:

Datum: 22.03.16

**Faculty of Linguistics and
Literary Sciences**

Prof.
John A. Bateman, PhD

Bibliothekstraße
GW 2, Raum A3480
28334 Bremen

Telefon (0421) 218 - 68120
Fax (0421) 218 - 98-68120
eMail bateman@
uni-bremen.de
www www-user.uni-
bremen.de/~bateman

I hereby confirm that Mr Thiago Pires was a guest PhD doctoral researcher in my working group at the University of Bremen from 1st April 2015 until 31st December 2015. During that time I supported him in his research concerning the relevance and impact of multimodality on automatic translation. He was fully integrated within my group for that time, participated in all of our activities and events on multimodality in general, and presented and discussed ongoing progress with myself and other members of the group. In the summer of 2015 he additionally presented interim research results in a poster session at our international Multimodality Conference. The results that were developed were of high relevance to our general research direction in multimodality and corpus-based approaches to multimodality in particular. The cooperation was therefore highly beneficial to both sides. We look forward to further possibilities for cooperation within Mr Pires in the future.

Yours faithfully,

Prof. John A. Bateman, PhD

Sekretariat
Sonja Kettler
Telefon (0421) 218 - 68053
Fax (0421) 218 - 98-68053
eMail skettler@
uni-bremen.de

ANEXO B – Autorização para uso do Google Tradutor

11/15/2016

Gmail - Your Google Cloud Platform Inquiry



Thiago Pires <thiagocomaga@gmail.com>

Your Google Cloud Platform Inquiry

3 messages

Andrews Moraes <andrewsmoraes@google.com>

Tue, Apr 19, 2016 at 10:42 AM

To: thiagocomaga@gmail.com

Hi Thiago,

Thanks for your inquiry with us here at the Google Cloud Platform Sales Team. I would really like to help you on this situation, so I'm going to try to point you to the best resources.

Unfortunately, there are no discounts available for educational organizations but there are Award Programs and Grants that you and your students can apply for. Your best resource for information is the [Research at Google: Award Programs](#) page.

Also, if you are interested in using Google App Engine, there is a free daily quota. For a breakdown of the free quota and pricing after the free quota is exceeded, please visit the [Google App Engine Pricing](#) page.

Lastly, I wanted to provide you with a link to a free trial so that you can try it now: [Google Cloud Console](#)

Best,
-Andy

Learn to build on the cloud today - [Google Cloud Platform Training](#)

Andrews Moraes | Google Cloud Platform | Global Market Development | andrewsmoraes@google.com


Thiago Pires <thiagocomaga@gmail.com>

Mon, May 23, 2016 at 12:47 PM

To: Andrews Moraes <andrewsmoraes@google.com>

Hi Andrews,

Do you know who should I get in contact to request an authorization to use Google translations for academic research purposes?

Best,

Thiago

[Quoted text hidden]

--

Prof. Thiago Blanch Pires
Departamento de Línguas Estrangeiras e Tradução - LET
Universidade de Brasília - UnB

Andrews Moraes <andrewsmoraes@google.com>

Mon, May 23, 2016 at 2:52 PM

To: Thiago Pires <thiagocomaga@gmail.com>

Google Translate API is open to the public, therefore there is no need for authorization requests. [Here are the Terms and Conditions](#)

-Andy

<https://mail.google.com/mail/u/0/?ui=2&ik=e3ee15ed6b&view=pt&q=andrews&qs=true&search=query&th=1542ec274eed6729&siml=1542ec274eed6729&sim...> 1/2

APÊNDICE A – Cronograma de atividades

2013	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Atividade1			x	x	x	x	x	x	x	x	x	x
Atividade2			x	x	x	x	x	x	x	x	x	x
Atividade3			x	x	x	x	x	x	x	x	x	x
<p>Atividade 1 – Aulas exigidas pelo PPGCINF/UnB;</p> <p>Atividade 2 – Levantamento, análise e discussão preliminar da bibliografia de trabalho (Arcabouço teórico-metodológico);</p> <p>Atividade 3 – Pesquisa exploratória, identificação e estudo de traduções automáticas a serem analisadas.</p>												

2014	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Atividade1			x	x	x	x	x					
Atividade2	x	x	x	x	x	x	x	x	x	x	x	x
Atividade3	x	x	x	x	x	x						
Atividade4				x	x	x	x	x	x	x	x	
Atividade5								x	x	x	x	x
Atividade6							x	x	x	x	x	x
<p>Atividade 1 – Aulas exigidas pelo PPGCINF/UnB;</p> <p>Atividade 2 – Levantamento, análise e discussão preliminar da bibliografia de trabalho (Arcabouço teórico-metodológico);</p> <p>Atividade 3 – Pesquisa exploratória, identificação e estudo de traduções automáticas a serem analisadas.</p>												

Atividade 4 – Organização do *International Workshop of Information Architecture and Multimodality*.
 Atividade 5 – Elaboração do relatório de qualificação.
 Atividade 6 – *Design* da base de dados da pesquisa.

2015	Jan	Fev	Mar	<u>Abr</u> ⁵⁸	<u>Mai</u>	<u>Jun</u>	<u>Jul</u>	<u>Ago</u>	<u>Set</u>	<u>Out</u>	<u>Nov</u>	<u>Dez</u>
Atividade1			x	x	x	x				x	x	x
Atividade2	X	x	x	X	x	x	x	x	x	x	x	x
Atividade3	x	x	x	X	X	X	x	x	x	x	x	x
Atividade4				X	x	x	x	x	x	x	X	
Atividade5									X	x	x	x
Atividade6										x	x	X
Atividade7											x	x

Atividade 1 – Aulas exigidas pelo grupo de pesquisa do exterior – *Universität Bremen*;
 Atividade 2 – Levantamento, análise, discussão avançada sobre a bibliografia de trabalho (Arcabouço teórico-metodológico);
 Atividade 3 – Estudo das ferramentas e técnicas usadas pelo grupo de pesquisa no exterior.
 Atividade 4 – Design e avaliação da base de dados da pesquisa.
 Atividade 5 – Identificação de relações intersemióticas.
 Atividade 6 – Categorização de relações intersemióticas.

⁵⁸ Os meses sublinhados e destacados representam o período de treinamento no exterior.

Atividade 7 – Tabulação dos achados em andamento.

2016	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Atividade1	x	x	x	x	x	x	x	x	x	x		
Atividade2	x	x										
Atividade3	x	x	x	x	x	x	x	x				
Atividade4			x									
Atividade5		x	x	x	x	x	x					
Atividade6				x	x	x	x	x	x	x	x	
Atividade7												x

Atividade 1 – Levantamento, análise, discussão avançada sobre a bibliografia de trabalho (Arcabouço teórico-metodológico);

Atividade 2 – Estudo das ferramentas e técnicas usadas pelo grupo de pesquisa no exterior.

Atividade 3 – Elaboração de uma proposta de análise de relações intersemióticas a partir dos achados do estudo piloto.

Atividade 4 – Qualificação da tese.

Atividade 5 – Sistematização dos achados.

Atividade 6 – Redação e revisão da tese.

Atividade 7 – Defesa da tese.

Fonte: Próprio autor⁵⁹.

⁵⁹ Tabela pode ser visualizada com mais qualidade a partir do arquivo excel encontrado dentro da pasta "corpus", disponível em

<https://www.dropbox.com/s/g7ky2traywqxbek/corpus.zip?dl=0>

APÊNDICE B – Tabela de classificação dos documentos multimodais

Nome do arquivo	Língua	Fonte	Formato do Meio	Tipo do conteúdo	Unidades básicas envolvidas	Tipos de incompatibilidades	Número de incompatibilidades	Outras relações	
Ways to Build a Bench_wikiHow_pt	Português	http://www.wikihow.com/Build-a-Bench	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos	a(3),tr(1),pt(1),rpd(1)	6	2
How to Make Coconut Flour With Flaked_pt	Português	http://www.wikihow.com/Make-Coconut-Flour-With-Flaked-Coconut	.png	Página Web	procedimental/explicativo	fotografias, legendas, subtítulos	a(1)	1	0
Como se livrar das pulgas na	Português	http://www.wikihow.com/Get-Rid-of-Fleas-Naturally	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos	a(4),rpd(1)	5	0
How to Get into Har_pt	Português	http://www.wikihow.com/Get-into-Harvard-Business-School	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos		0	0
How to Fix a Squeaking Bed_pt	Português	http://www.wikihow.com/Fix-a-Squeaking-Bed-Frame	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos	a(3),rpd(1)	4	0
How to Create Bus_pt	Português	http://www.wikihow.com/Create-Business-Letter-White-Space	.png	Página Web	procedimental/explicativo	fotografias, legendas, subtítulos		0	0
Como se tornar um acadêmico_pt	Português	http://www.wikihow.com/Become-an-Academic	.png	Página Web	procedimental/explicativo	desenhos, legendas, subtítulos		0	0
4 Ways to Clarify Butter_pt	Português	http://www.wikihow.com/Clarify-Butter	.png	Página Web	procedimental/explicativo	fotografias, legendas, subtítulos	a(5)	5	0
25 anos após a independência_ng_pt	Português	http://www.nationalgeographic.com/photography/proof/2016/09/taj	.pdf	Página Web	informacional/picture story	fotografias, legendas, título, subtítulos	ai(1)	0	0
Lion_ng_pt	Português	http://kids.nationalgeographic.com/animals/lion/	.pdf	Página Web	informacional/educacional	fotografias, desenhos, título, subtítulo, legend	ai(3)	3	0
Brazil dam burst engul_pt	Português	http://www.bbc.com/news/world-latin-america-34742272	.pdf	Página Web	informacional	fotografias, vídeos, legendas, tópico, subtópico		0	0
Brexit_David Cameron to quit_pt	Português	http://www.bbc.com/news/uk-politics-36615028	.pdf	Página Web	informacional	fotografias, vídeos, mapa, legendas, tópico, si	ai(1)	1	0
Islâmico apreende	Português	http://www.bbc.com/news/world-middle-east-32820857	.pdf	Página Web	informacional	fotografias, vídeos, mapa, legendas, tópico, si	ai(1)	1	0
Sinai plane crash_Russia sus_pt	Português	http://www.bbc.com/news/world-middle-east-34687990	.pdf	Página Web	informacional	fotografias, mapa, legendas, título, subtítulos		0	0
The beauty of the Montreal Met_pt	Português	http://www.bbc.com/news/in-pictures-34679942	.pdf	Página Web	informacional/picture story	fotografias, legendas, título, subtítulos		0	0
US rejects Keystone XL pip_pt	Português	http://www.bbc.com/news/world-us-canada-34751370	.pdf	Página Web	informacional	fotografias, mapa, legendas, título, subtítulos		0	0
19 Secrets Librarians Will Nev_pt	Português	https://www.buzzfeed.com/farahpenn/secrets	.png	Página Web	informacional/entretimento	fotografias, desenhos, título, subtítulo, legend	ai(4)	4	0
These Paralympians Ran_pt	Português	https://www.buzzfeed.com/javiermoreno/these-paralympians-ran-the	.png	Página Web	informacional/entretimento	fotografias, vídeos, legendas, tópico, subtópico		0	0
creative_blog_google_results_en	Português/Inglês	https://www.trendhunter.com/trends/news/en/news-room/20141202FG82	.png	Infográfico	procedimental/informacional	desenhos, legendas, títulos, e subtítulos	a(1)	1	0
europarl.europa.eu_pt	Português	http://www.europarl.europa.eu/news/en/news-room/20141202FG82	.png	Infográfico	informacional	gráfico, legenda, título, e subtítulo		0	0
Museum development_pt	Português/Inglês	http://www.gbstudioart.co.nz/blog/410257	.png	Infográfico	informacional	desenho, título, subtítulo, balões de fala, seta		0	0
These Are The Foods We'd Lose If Honey_pt	Português/Inglês	http://www.huffingtonpost.com/entry/honeybees-food-colony-collapse	.png	Infográfico	informacional	desenho, título, subtítulo		0	0
tooth_truths_en	Português/Inglês	https://www.televox.com/10-truths-about-the-tooth-infográfico/	.png	Infográfico	informacional/explicativo	desenho, legenda, título		0	0
Airbus-AC_A321_01_May_2015	Português/Inglês	http://www.docfoc.com/airbus-aca32101may2015	.pdf	Manual	procedimental/explicativo	desenho, título, subtítulo, gráficos, tabelas,		0	0
eosr3-eos1100d-im2-c-en	Português/Inglês	http://gdlp01.c-wss.com/gds/0/0300004730/02/eosr3-eos1100d-im2	.pdf	Manual	procedimental/explicativo	desenhos, fotografias, títulos, subtítulos		0	0
HERO3_UM_Black_ENG_REV0_WEB	Português/Inglês	http://www.lightson.com/rental/owners_manuals/HERO3_UM_Black	.pdf	Manual	procedimental/explicativo	desenhos, ícones, tabelas, títulos, subtítulos		0	0
WIF_SM-T800_TAB_S_EN_UM_KK_FINAL	Português/Inglês	http://downloadcenter.samsung.com/content/UM/201512/20151208	.pdf	Manual	procedimental/explicativo	desenhos, ícones, tabelas, títulos, subtítulos		0	0
Taboada_Habel_Discourse_Studies	Português/Inglês	http://dis.sagepub.com/content/15/1/65	.pdf	Artigo acadêmico	informacional/explicativo	fotografias, gráficos, tabelas, títulos, subtítulo		0	0
Um quadro para a análise multimodal de _ - http_www.tandfonline.com/doi/ful/10.1080/10350330.2014.100217	Português	http://www.tandfonline.com/doi/ful/10.1080/10350330.2014.100217	.pdf	Artigo acadêmico	informacional/explicativo	tabelas, fotografias, captura de tela, legendas,		0	0
Romero-Fresco article - http_www.jostrans.org/issue20/art_romero.php	Português	http://www.jostrans.org/issue20/art_romero.php	.pdf	Artigo acadêmico	informacional/explicativo	capturas de telas, fotografias, tabelas, títulos, a(1)		0	0

Fonte: Próprio autor⁶⁰.

⁶⁰ Tabela pode ser visualizada com mais qualidade a partir do arquivo excel encontrado dentro da pasta "corpus", disponível em

<https://www.dropbox.com/s/g7ky2traywqxbek/corpus.zip?dl=0>