ANAIS DO X ENCONTRO DE LINGUÍSTICA DE CORPUS

ASPECTOS METODOLÓGICOS DOS ESTUDOS DE CORPORA

Deise Prina Dutra
Heliana Mello
Organizadoras

ANAIS DO X ENCONTRO DE LINGUÍSTICA DE CORPUS

ASPECTOS METODOLÓGICOS DOS ESTUDOS DE CORPORA

Belo Horizonte
Faculdade de Letras da UFMG
2012

Universidade Federal de Minas Gerais
Reitor: *Clélio Campolina Diniz*
Vice-Reitora: *Rocksane de Carvalho Norton*

Faculdade de Letras
Diretor: Luiz Francisco Dias
Vice-Diretora: Sandra Maria Gualberto Braga Bianchet

Projeto Gráfico e editoração: Marco Antônio Durães e Alda Lopes
Capa: Marco Antônio Durães e Alda Lopes, baseado em cartaz elaborado pelo CEDECOM.

Ficha catalográfica elaborada pelos Bibliotecários da Biblioteca FALE/UFMG

E56a

Encontro de Linguística de Corpus (10. : 2012 : Belo Horizonte, MG.
Anais do X Encontro de Linguística de Corpus : aspectos metodológicos dos esudos de corpora / Deise Prina Dutra, Heliana Ribeiro de Mello (organizadoras). – Belo Horizonte : Faculdade de Letras da UFMG, 2012.
380 p.: il.

Inclui bibliografias.

ISBN: 978-85-7758-153-5

1. Linguística de corpus - Congressos. 2. Semântica - Congressos. I. Dutra, Deise Prina. II. Mello, Heliana Ribeiro de. III. Universidade Federal de Minas Gerais. Faculdade de Letras. IV. Título.

CDD : 410.285

Sumário

Apresentação

É com muita satisfação que apresentamos os Anais do X Encontro de Linguística de Corpus. O X ELC foi realizado na Faculdade de Letras da Universidade Federal de Minas Gerais em 11 e 12 de novembro de 2011. O evento contou com a presença de convidados nacionais e internacionais que proferiram plenárias e participaram em uma mesa redonda, assim como de pesquisadores que apresentaram seus trabalhos nas modalidades de comunicação oral, pôsteres e sessão de trabalhos em andamento.

A nossa avaliação como organizadoras é de que o evento foi um grande sucesso e representou mais um passo na consolidação do campo de estudos da Linguística de Corpus no Brasil. Assim, nesta publicação que ora lhes apresentamos, constam trabalhos que versam sobre pesquisas realizadas através de metodologia de análise de corpora. Os trabalhos foram agrupados em cinco grupos temáticos: Panorama; Linguística de Corpus e Semântica; Frames, Construções e Anotação de Corpora; Gramática e Estrutura Informacional; Tradução, Comparação Interlinguística e Ensino de Línguas.

Deixamos aqui o nosso agradecimento aos colegas que participaram do X ELC, em especial àqueles que submeteram seus trabalhos à publicação. Para os leitores que já são pesquisadores do campo da Linguística de Corpus esperamos estar oferecendo uma leitura que venha a auxiliá-los em seus próprios trabalhos; para aqueles que estão sendo apresentados à área pela primeira vez, esperamos que sejam capturados pela vasta gama de recursos e possibilidades que essa área de pesquisa nos permite. A todos desejamos uma ótima leitura!

Deise Prina Dutra & Heliana Mello

I PANORAMA

Pela visibilidade da Linguística de Corpus em trabalhos acadêmicos

Daniel Alves[1]
Stella E. O. Tagnin[2]

RESUMO: A partir de uma amostra de 84 trabalhos acadêmicos que adotam suporte teórico-metodológico da Linguística de Corpus, foi analisada a visibilidade da Linguística de Corpus em dados que compõem as informações catalográficas (como títulos, palavras-chave e resumos). Constatou-se que faltam, em alguns trabalhos, referências claras quanto à utilização da Linguística de Corpus – o que pode ter impactos sobre a recuperação de informações da área. Neste artigo, são expostos os passos que levaram a essa constatação e são feitas sugestões de definição de títulos e palavras-chave, com vistas a aumentar a visibilidade da Linguística de Corpus em trabalhos acadêmicos e, consequentemente, otimizar a recuperação de informações por parte de interessados na área.

PALAVRAS-CHAVE: Linguística de Corpus, Análise bibliométrica de trabalhos acadêmicos, Visibilidade da Linguística de Corpus em textos acadêmicos.

ABSTRACT: An analysis of 84 academic texts on Corpus Linguistics was carried out and revealed a lack of clearness in providing information on the affiliation of these texts with Corpus Linguistics – a problem which can have a direct impact on the retrieval of information in the area. This chapter intends to suggest an initiative to standardize the vocabulary used for the keywords, which should, we hope, increase the visibility of Corpus Linguistics in academic texts, optimizing the retrieval of information.

KEYWORDS: Corpus Linguistics; Bibliometric analysis of academic texts; visibility of Corpus Linguistics in academic texts.

[1] Daniel Alves é professor do curso de Bacharelado em Tradução da Universidade Federal da Paraíba, doutorando em Estudos da Tradução pela Universidade Federal de Santa Catarina - daniel.alves.ufpb@gmail.com.

[2] Stella E. O. Tagnin é professora associada da Universidade de São Paulo e doutora em Estudos Linguísticos e Literários em Inglês pela Universidade de São Paulo - seotagni@usp.br.

1 Introdução

Este artigo tem como objetivo propor uma metodologia de identificação que aumente a visibilidade da Linguística de Corpus em trabalhos acadêmicos, facilitando, assim, a recuperação de informações na área.

Considerando o atual contexto mundial de profusão de informações e a crescente dificuldade em se recuperar informações precisas nos momentos adequados, o trabalho aqui apresentado busca investigar a facilidade de recuperar trabalhos acadêmicos que adotam princípios da Linguística de Corpus em suas composições. Para tanto, é analisada uma amostra de 84 trabalhos desenvolvidos na área e é verificado se tais trabalhos apresentam – com clareza em seus títulos, palavras-chave e resumos – informações que facilitem o levantamento bibliográfico.

O artigo se divide em cinco seções além desta introdução, a saber: a) A recuperação de informações no mundo contemporâneo – em que se discutem o crescimento da produção de informações e a preocupação de diferentes áreas do conhecimento em adotar mecanismos que possibilitem recuperar a informação necessária no momento oportuno; b) A identificação da Linguística de Corpus em trabalhos acadêmicos – em que são analisadas informações catalográficas (como títulos, palavras-chave e resumos) de uma amostra de 84 trabalhos acadêmicos, buscando identificar se fica claro – para um(a) leitor(a) desses trabalhos – a adoção da Linguística de Corpus como referencial teórico-metodológico; c) Áreas de pesquisa que utilizam suporte teórico-metodológico da Linguística de Corpus – em que, a partir da amostra analisada, é traçado um panorama geral da utilização da Linguística de Corpus em pesquisas acadêmicas, contemplando diferentes áreas de investigação, idiomas e gêneros textuais mais investigados; d) Propostas para aumentar a visibilidade da Linguística de Corpus – em que são sugeridos um vocabulário padrão e uma metodologia no intuito de aumentar a visibilidade da Linguística de Corpus em trabalhos acadêmicos; e e) Considerações finais – em que são retomados os resultados da pesquisa aqui apresentada e feitos os encaminhamentos finais deste trabalho.

A próxima seção apresenta considerações sobre o contexto contemporâneo de produção de informações e sobre a preocupação de diferentes áreas do conhecimento em desenvolver métodos para recuperar informações de forma eficiente.

2 A recuperação de informações no mundo contemporâneo

O volume de informações produzido pela humanidade cresce a um ritmo cada vez mais acelerado. Segundo Hilbert & López (2011, p. 62), apenas em 2007, a quantidade total de informações produzida no mundo foi de 295 exabytes (ou 295 trilhões de megabytes) – contra 2,6 exabytes em 1986; 15,8 em 1993; e 54,5 em 2000. Para se ter um parâmetro físico sobre o que corresponde a tal quantidade de informação, Hilbert & López (2011:62) apontam que se todas as informações produzidas em 2007 fossem armazenadas em CD-ROMs de 1,2mm de espessura e se esses CDs fossem empilhados, a pilha produzida teria 1,25 vezes a distância da terra à lua.

Nesse contexto de produção cada vez maior de volumes de dados, é possível identificar, nas diferentes áreas do conhecimento, a preocupação com a recuperação de informações. Encontrar a informação precisa no momento adequado tornou-se um tema que ultrapassa as fronteiras acadêmicas que separam as áreas da ciência. A ciência da informação e a ciência da computação são duas dessas áreas – para limitar o número de exemplos – que se preocupam em buscar estratégias para recuperar informações de forma eficiente.

No âmbito da Ciência da Informação, por exemplo, uma das abordagens utilizadas para a recuperação de informações em grandes bases de dados é a elaboração de taxonomias (ou classificações sistemáticas de dados). Segundo Campos & Gomes (2008), as taxonomias têm sido cada vez mais empregadas em portais corporativos e em bibliotecas digitais "por permitir[em] acesso através de uma navegação em que os termos se apresentam de forma lógica".

Já no âmbito da Ciência da Computação, técnicas de mineração de dados têm sido utilizadas para descobrir padrões em bancos de dados, recuperando informações com técnicas matemáticas que evitam limitações e tendenciosidades humanas, como aponta Braga (2005:11). Trata-se de um "conjunto de técnicas para descrição e predição a partir de grandes massas de dados" (Braga, 2005, p. 12) que busca levantar informações precisas, completas e relevantes.

O público que estuda a Linguística de Corpus, no entanto, nem sempre tem controle sobre a classificação das informações disponibilizadas em bancos de dados, como têm os(as) profissionais da Ciência da Informação, e/ou algoritmos avançados de levantamento de padrões, como os(as) da Ciência da Computação. Trata-se, aqui, de um público que deseja recuperar informações, contando com recursos disponíveis apenas para usuários.

Neste artigo, pretendemos nos dirigir aos(às) pesquisadores(as) da Linguística de Corpus, propondo um vocabulário-padrão, baseado em critérios previamente estabelecidos, para otimizar a recuperação de trabalhos sobre Linguística de Corpus e/ou que adotam princípios da Linguística de Corpus em suas metodologias. Com isso, pretendemos aumentar a visibilidade dos trabalhos relacionados à Linguística de Corpus e, assim, facilitar os trabalhos de pesquisa dos usuários mencionados no parágrafo anterior.

Partimos, como mencionado anteriormente, da análise de uma amostra do atual quadro das pesquisas relativas à Linguística de Corpus no Brasil. A partir de um levantamento sobre as publicações na área, apontamos as dificuldades encontradas ao se levantar a respectiva bibliografia. Na próxima seção, será apresentado como, atualmente, são identificados os trabalhos acadêmicos que investigam a Linguística de Corpus no Brasil.

3 A identificação da linguística de corpus em trabalhos acadêmicos

Inicialmente, foi feito o levantamento de trabalhos orientados por pesquisadores(as) brasileiros(as) que se dedicam a estudos que contam com a Linguística de Corpus como suporte teórico-metodológico. Desse levantamento, resultaram 84 trabalhos, que, para os fins desta análise, serão tratados como uma amostragem dos estudos baseados na Linguística de Corpus atualmente desenvolvidos no Brasil.

Como anteriormente mencionado, a investigação dessa amostra teve como objetivo identificar se, na área da Linguística de Corpus, era clara a identificação da malha teórica a partir dos títulos, palavras-chave e outras informações catalográficas.

Os trabalhos que constituem a amostra das pesquisas atualmente desenvolvidas nessa área no Brasil foram organizados em uma planilha eletrônica, com informações como Filiação, Nível da Pesquisa, Pesquisador(a), Orientador(a), Título do Trabalho e outras. Com isso, foi possível sistematizar os dados sobre as pesquisas e fazer um diagnóstico sobre a visibilidade que tais pesquisas teriam em uma busca relacionada à Linguística de Corpus.

O primeiro ponto que se destacou, nesse diagnóstico, foi o a dificuldade de identificação, não apenas do recurso à Linguística de

Corpus, mas também da própria área em que se inserem os trabalhos. Dos 84 trabalhos analisados, 44 (o equivalente a 52%) não fazem referência à Linguística de Corpus em seus títulos. Os títulos apresentados a seguir, por exemplo, são alguns dos casos que mostram esse tipo de problema:

- Metáforas dos líderes
- A desmetaforização como hipótese produtiva para a modelagem do processo tradutório
- A transferência em estrutura argumental português-inglês
- Reasons for the English vocabulary increase in the 16th century
- O gênero notícia científica no jornal televisivo brasileiro

A partir dos exemplos apresentados, nota-se a dificuldade em identificar a área de concentração das pesquisas em tela. Em casos como esses, não são facilmente identificáveis a malha teórica, as metodologias de investigação nem mesmo as linhas de investigação da pesquisa.

Apesar dos problemas levantados nos 84 trabalhos que compuseram a amostra analisada, foi possível, a partir deles, atestar a característica da Linguística de Corpus de ser uma ferramenta poderosa que oferece suporte teórico-metodológico a diversas áreas da pesquisa linguística. Na próxima seção, apresentaremos essa característica de pluralidade, mostrando áreas de pesquisa que se embasam na Linguística de Corpus, seus principais focos e características mais salientes.

4 Áreas de pesquisa que utilizam suporte teórico-metodológico da linguística de corpus

A amostra de trabalhos que analisamos evidenciou a pluralidade de aplicações da Linguística de Corpus como ferramenta teórico-metodológica para pesquisa linguística. Nesta seção, será apresentada uma classificação das pesquisas analisadas neste trabalho – considerando a área de concentração e os principais interesses de pesquisa.

Identificarmos cinco áreas em que a Linguística de Corpus está presente, a saber: a) tradução; b) descrição da linguagem; c) ensino;

d) terminologia e; e) Processamento de Linguagem Natural (PLN). O gráfico abaixo apresenta a distribuição dos trabalhos analisados em cada uma dessas cinco áreas.

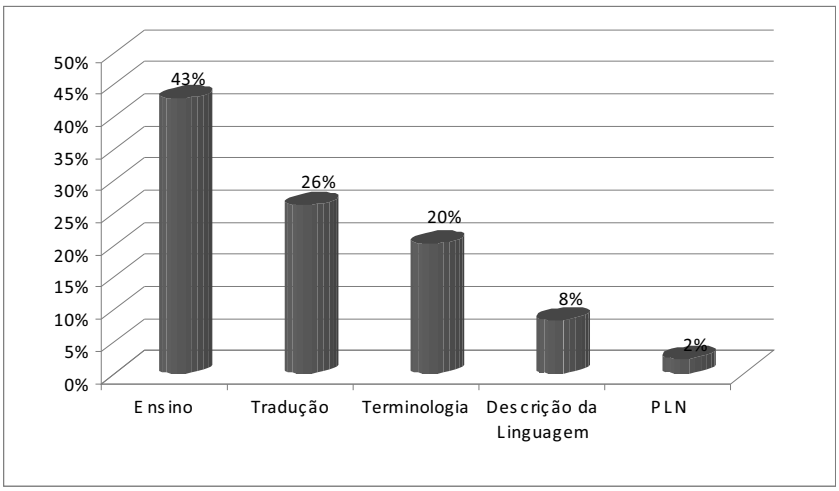


GRÁFICO 1
Áreas em que a Linguística de Corpus está presente

Como é possível ver, no GRÁFICO 1, dos trabalhos que constituíram a amostra aqui analisada, a maior parte se insere nas áreas de Ensino (43%) e Estudos da Tradução (26%), seguidos por Terminologia (20%), Descrição da Linguagem (8%) e Processamento de Linguagem Natural (2%). Chama a atenção, nessa distribuição, a baixa participação dos trabalhos de Processamento de Linguagem Natural. Considerando-se que os trabalhos nessa área, em grande parte, tendem a recorrer a corpora para desenvolver suas ferramentas, esperava-se maior participação no cômputo geral.

A classificação apresentada, no entanto, não foi pacífica. Como já mencionado, houve dificuldades para classificar alguns dos trabalhos nas áreas identificadas por falta de clareza na localização teórica desses trabalhos. No QUADRO 1, a seguir, apresentamos alguns exemplos desses títulos e as dúvidas que tivemos ao tentar classificá-los:

QUADRO 1
Exemplos de títulos de trabalhos que suscitaram dúvidas de classificação

Títulos dos trabalhos	**Possíveis classificações**
• O uso da metalinguagem no discurso em sala de aula do professor de língua estrangeira.	Descrição de linguagem ou Ensino
• The Place of Grammar in Teacher Talk	Descrição de linguagem ou Ensino
• Organização temática em corpus paralelo bilíngüe no registro ficcional.	Descrição de linguagem ou Tradução

Em casos como os apresentados no QUADRO 1, em que havia dúvidas quanto à classificação, foram buscadas características que pudessem associar o trabalho a uma área de pesquisa mais específica. Na última linha do QUADRO 1, por exemplo, a opção da classificação foi pelo rótulo 'Tradução' por se tratar de um 'corpus paralelo'. É importante ressaltar, no entanto, que classificações são atividades eminentemente epistemológicas, sempre sujeitas a críticas também epistemológicas. As classificações que adotamos neste trabalho tiveram como objetivo traçar um panorama bastante genérico da presença da Linguística de Corpus em trabalhos acadêmicos e não se propuseram, de forma alguma, a fechar o debate sobre o tema.

As seções a seguir apresentam informações sobre o panorama que traçamos, indicando, nas áreas de Ensino, Tradução, Terminologia, Descrição da Linguagem, e PLN, as principais línguas investigadas, gêneros textuais trabalhados e alguns focos de interesse.

a. Ensino

Em relação às pesquisas que investigam o ensino utilizando ferramentas de corpora, o perfil aqui traçado identificou um forte interesse pela investigação de usos de ferramentas de corpora no ensino de inglês (espanhol e português para estrangeiros também têm espaço, mas com menor representatividade), como mostra o gráfico a seguir:

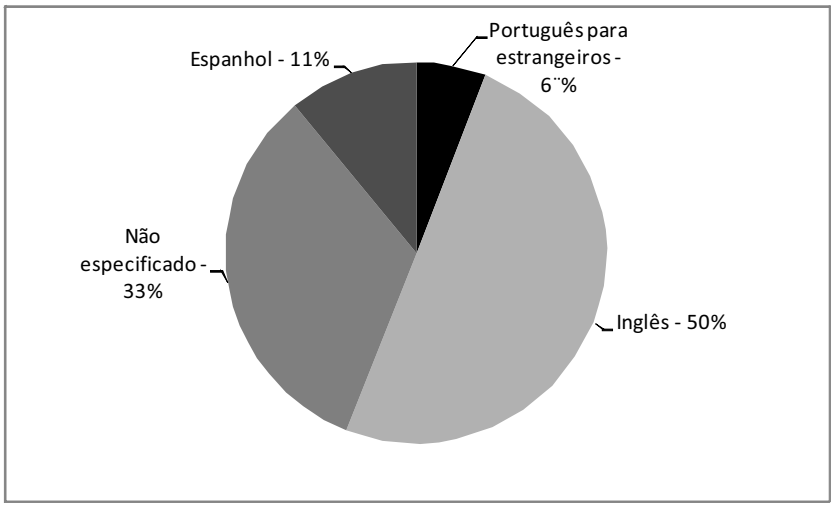


GRÁFICO 2
Distribuição das pesquisas voltadas para o ensino de língua estrangeira

Metade das pesquisas que compõem a amostra dedica-se ao recurso a corpora no ensino da língua inglesa. Mais uma vez, é alto o percentual de pesquisas que não deixam claro, em seus títulos, as línguas (ou os pares linguísticos) trabalhadas. O QUADRO 2, a seguir, apresenta alguns exemplos dos títulos de pesquisas que investigam o uso de corpora aplicados ao ensino:

QUADRO 2
A recuperação dos idiomas nos títulos dos trabalhos

Títulos dos trabalhos	**Informações sobre os pares linguísticos trabalhados**
• Preparação de materiais para e-classes com corpora eletrônicos	Não especificado
• Usando Erros para Produzir Acertos em Cursos On-line Um Estudo Baseado em Corpus de Aprendiz	Não especificado
• Atividades de ensino para espanhol com Lingüística de Corpus	Espanhol
• A linguagem de Role Playing Games Digitais e o Ensino de Inglês	Inglês

Como se pode ver nos casos apresentados no QUADRO 2, os dois primeiros trabalhos não dão indícios sobre a(s) língua(s) enfocada(s) nas análises, ao passo que os dois últimos o fazem. Destaca-se, no quadro, o terceiro exemplo, em que são feitas referências explícitas tanto ao idioma trabalhado quanto à Linguística de Corpus – uma característica de clareza que facilita trabalhos de classificação (como o aqui realizado) e de recuperação desse artigo por estudiosos(as) interessados(as) no tema.

O campo do Ensino é bastante amplo e variado, como se observa no GRÁFICO 3, a seguir:

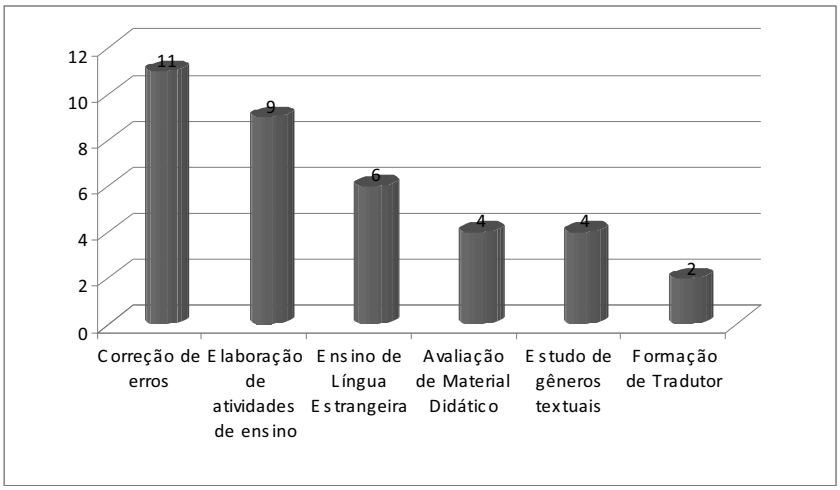


GRÁFICO 3
Principais focos de atenção das pesquisas voltadas
ao Ensino que utilizam corpora

Como se pode ver no GRÁFICO 3, a correção de erros lidera as abordagens relacionadas ao ensino, seguida da elaboração de atividades de ensino. O ensino de língua estrangeira, a avaliação de material didático e o estudo de gêneros textuais representam pouco menos da metade da totalidade dos trabalhos. Note-se, no entanto, que o ensino de tradução, visando a formação do tradutor, também está aqui representada, embora modestamente.

b. Tradução

Como apontado no GRÁFICO 1, uma das áreas mais produtivas (segundo nossa amostra) que se amparam na Linguística de Corpus foi a de Estudos da Tradução. De forma a traçar um panorama dessas pesquisas, foram investigados os pares linguísticos e os gêneros textuais analisados nessas pesquisas.

A partir dos títulos, foram feitas tentativas de recuperar os pares linguísticos investigados. Essa análise revelou que Inglês e Português são as línguas mais investigadas, mas que grande parte das pesquisas (44%) não dá indícios, direta ou indiretamente, das línguas trabalhadas, como mostram o GRÁFICO 4 e o QUADRO 3, a seguir:

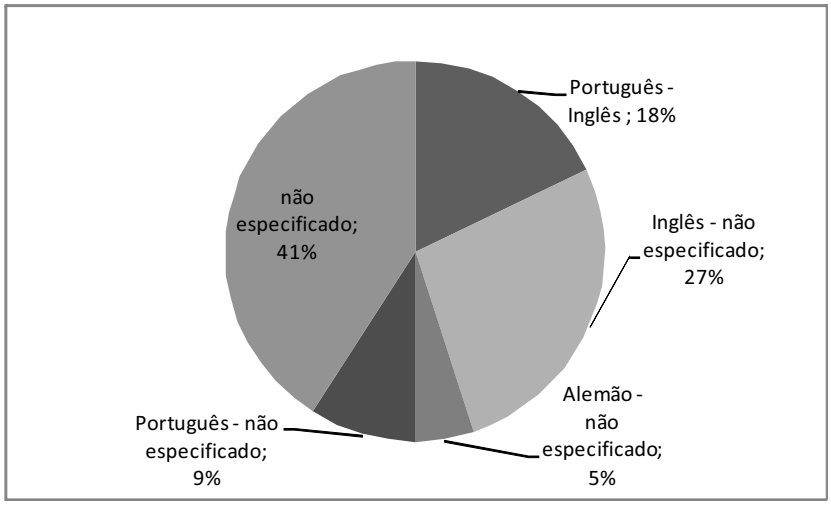


GRÁFICO 4
Pares linguísticos trabalhados nas pesquisas em Estudos da Tradução que utilizam a Linguística de Corpus

O quadro a seguir exemplifica a recuperação dos pares linguísticos a partir dos títulos dos trabalhos:

QUADRO 3
A recuperação dos pares linguísticos em títulos de trabalhos em Estudos da Tradução

Títulos dos trabalhos	**Informações sobre os pares linguísticos trabalhados**
• Análise de tradução bem-sucedida com Linguística de Corpus	Não especificado
• Terminologia e Tradução - a interface que precisa ser trabalhada	Não especificado
• Estudo comparativo de duas traduções brasileiras da peça *Pygmalion* de Bernard Shaw: Desafios do dialeto Cockney	Português-Inglês
• *Homepages* institucionais em português e suas versões para o inglês: Uma análise baseada em corpus de aspectos lexicais e discursivos	Português-Inglês

Nos dois primeiros casos apresentados no QUADRO 3, os títulos não fazem referência aos idiomas trabalhados na pesquisa, tampouco fornecem quaisquer indícios que permitam recuperar essa informação. Já no terceiro caso, há menção explícita do autor da obra original (Bernard Shaw, autor de textos em língua inglesa), além de fazer referência explícita à tradução brasileira da peça trabalhada. Já no quarto caso, a recuperação da informação sobre o par linguístico abordado não apresenta dificuldades, considerando que são feitas referências explícitas às línguas portuguesa e inglesa.

Sobre os gêneros textuais investigados, foi possível notar que a maior parte das pesquisas direciona-se para textos literários (considerando Romances, Literatura Infanto-Juvenil e Teatro), como mostra o gráfico abaixo. Novamente, a partir dos títulos dos trabalhos não foi possível, em todos os casos, determinar os gêneros textuais trabalhados, como mostram o GRÁFICO 5 e o QUADRO 4 a seguir:

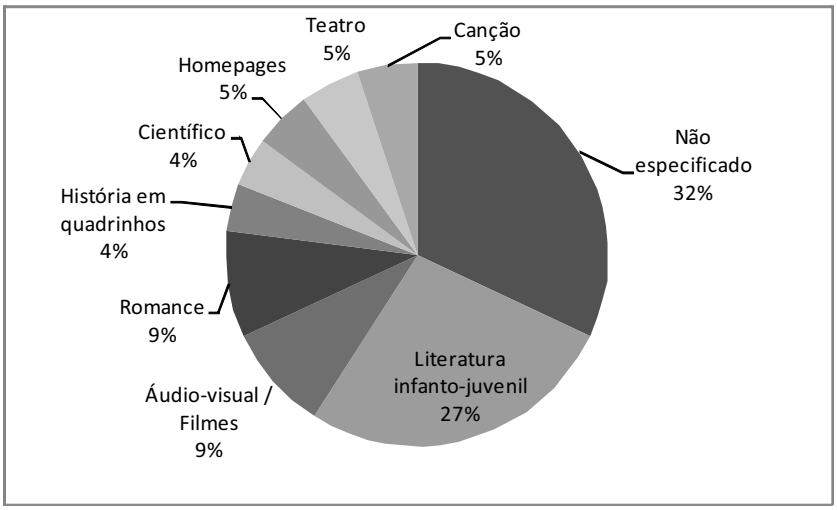


GRÁFICO 5
Gêneros textuais mais trabalhados nas pesquisas em Estudos da Tradução

O quadro a seguir exemplifica a recuperação dos gêneros textuais a partir dos títulos dos trabalhos:

QUADRO 4
A recuperação de informação dos gêneros textuais trabalhados a partir dos títulos dos trabalhos

Títulos dos trabalhos	**Informações sobre os gêneros textuais trabalhados**
• Análise de tradução bem-sucedida com Linguística de Corpus	Não especificado
• Terminologia e Tradução - a interface que precisa ser trabalhada	Não especificado
• Façanhas e limitações da tradução de canções: Um estudo a partir de versões de Garota de Ipanema em cinco idiomas	Canção
• Estudos da Tradução sobre Literatura Infantil: a retextualização de Flicts em Língua Inglesa	Literatura infanto-juvenil

Como nos casos anteriores, o QUADRO 4 exemplifica dois trabalhos que não especificam o gênero abordado e dois trabalhos em que essa informação é explicitada.

c. Terminologia

Para traçar um perfil das pesquisas em Terminologia, foram analisadas as línguas e os domínios abordados nas pesquisas terminológicas. Em relação às línguas investigadas, mais uma vez notou-se a presença marcante do inglês, seja no par português-inglês, seja em trabalhos monolíngues. O gráfico a seguir mostra a distribuição das pesquisas por língua:

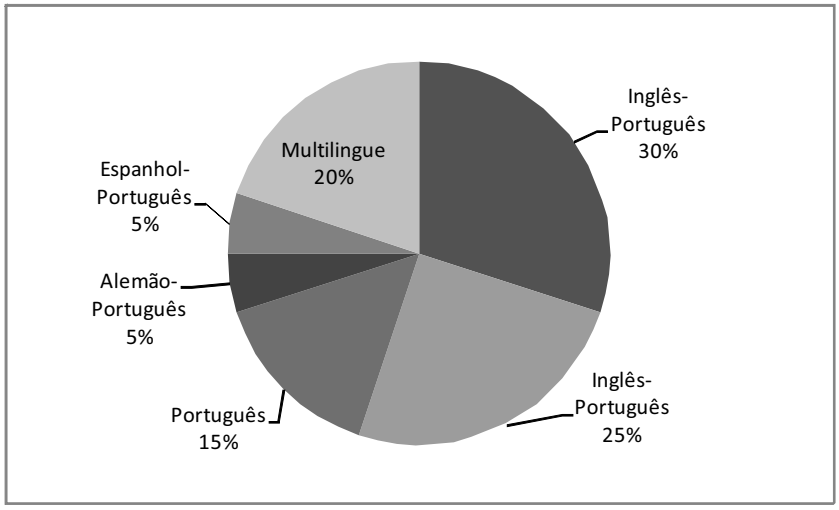


GRÁFICO 6
Distribuição das pesquisas em Terminologia,
que utilizam corpora, por língua

Como se pode ver no GRÁFICO 6, inglês e português são as línguas mais enfocadas nas pesquisas em corpora voltadas para a área de Terminologia – fenômeno semelhante ao observado nas pesquisas em Tradução e Descrição da Linguagem, como anteriormente apontado.

Os domínios abordados nas pesquisas terminológicas são bastante variados, encabeçados pela Medicina, seguida pela Economia e pelas Ciências Humanas. Outras áreas contempladas são:

jurídica, juramentada, nanociência, culinária, turismo, bio-combustível, técnica-científica, escolar, turismo e hotelaria. Os tópicos abordados são igualmente variados, conforme se vê no gráfico a seguir, onde se salienta a área emergente da Interpretação:

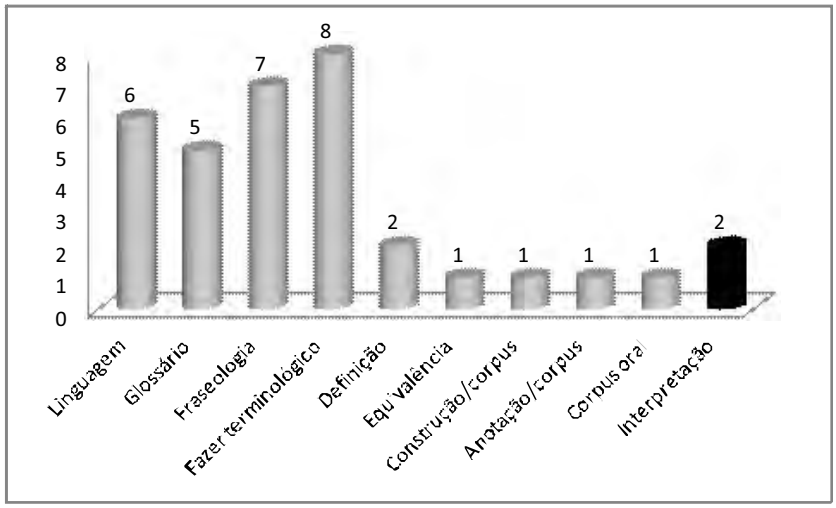


GRÁFICO 7
Principais focos de interesse das pesquisas em Terminologia

Como se pode ver no GRÁFICO 7, o maior foco de interesse das pesquisas nesta área está sobre o fazer terminológico, que contempla reflexões sobre o processo de elaboração de produtos terminológicos.

O QUADRO 5, a seguir, traz alguns exemplos de títulos de trabalhos que se inserem na área de Terminologia, mostrando que, também nessa área, a classificação nem sempre foi pacífica.

QUADRO 5
A recuperação de informação sobre a área de Terminologia
a partir dos títulos dos trabalhos

Títulos dos trabalhos	**Informações sobre o foco da investigação terminológica**
• Protocolos de anotação de corpora em MMAX2	Não especificado
• Interpretação Simultânea: a Linguística de Corpus na preparação do intérprete	Não especificado
• A compilação de um glossário bilíngue de colocações, na área de negócios, baseado em corpus comparável	Explícito (glossário)
• Terminologia em português da Nanociência e Nanotecnologia: confecção de corpus e de base terminológica	Explícito (Terminologia; base terminológica)

Os dois primeiros exemplos apresentados no QUADRO 5 mostram casos em que os títulos dos trabalhos não deixam claro que se inserem na área de investigação de Terminologia. O segundo exemplo é bastante emblemático disso: embora aborde a terminologia ao investigar a Interpretação Simultânea, o título do trabalho não deixa isso claro para os(as) interessados(as). Já no terceiro e no quarto exemplos, há casos em que fica explícita (pelo uso de termos como glossário; terminologia; base terminológica) a abordagem teórica das investigações apresentadas.

d. Descrição da Linguagem e PLN

O baixo número de pesquisas relativas à Descrição da Linguagem e ao Processamento de Linguagem Natural (respectivamente sete e dois trabalhos) em nossa amostra não nos permite fazer maiores generalizações. Foi possível, no entanto, perceber que – assim como observado em outras áreas de pesquisa – a maior parte dos trabalhos também não especifica as línguas investigadas, como mostra o GRÁFICO 8, a seguir:

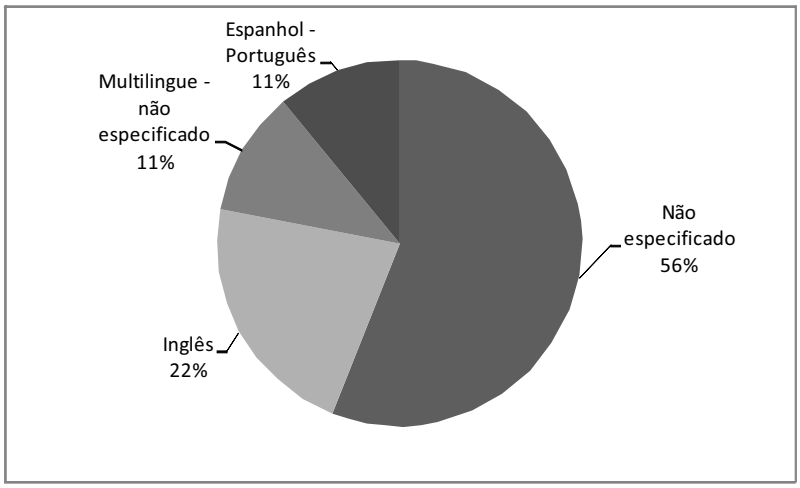


GRÁFICO 8
Línguas em que se concentram as pesquisas em
Descrição da Linguagem e PLN

Como se pode ver no GRÁFICO 8, a maior parte das pesquisas em Descrição da Linguagem e PLN (56%), não dá indícios, em seus títulos, das línguas sobre as quais se desenvolvem. No entanto, como anteriormente dito, o baixo número de pesquisas em nossa amostra não permite fazer maiores generalizações. O quadro a seguir traz exemplos de títulos de pesquisas em Descrição da Linguagem, enfatizando a dificuldade encontrada em identificar o par linguístico trabalhado:

QUADRO 6
A recuperação de informações sobre as línguas investigadas
em trabalhos de Descrição da Linguagem e LN

Títulos dos trabalhos	**Informações sobre as línguas trabalhadas**
• O corpus do jornal: variação linguística, gêneros e dimensões da imprensa diária escrita	Não especificado
• Dimensões de variação em apresentações orais e artigos acadêmicos da área médica: Um estudo baseado em corpus	Não especificado
• Perfil léxico-gramatical dos textos das provas de inglês do vestibular da FUVEST: Um estudo baseado em corpus	Inglês
• Falsos amigos no Espanhol para alunos brasileiros	Espanhol

5 Propostas para aumentar a visibilidade da linguística de corpus

Nas seções anteriores foi exposta a atual dificuldade de recuperar informações, dado o contexto de produção de informações em que a humanidade se encontra, e foi apresentado o caráter plural da Linguística de Corpus – que funciona como suporte teórico-metodológico de diferentes áreas da pesquisa linguística (abrangendo ensino, tradução, terminologia, descrição de linguagem e processamento de linguagem natural).

Nesta seção, pretendemos sugerir, aos(às) pesquisadores(as) que utilizam a Linguística de Corpus como suporte teórico-metodológico, um vocabulário-padrão e uma metodologia de hierarquização de palavras-chave que possa facilitar o trabalho de recuperação de informações por parte de pesquisadores(as) interessados(as) na área. Cabe aqui diferenciar, de um lado, os trabalhos que têm a Linguística de Corpus no cerne de suas investigações e de suas perguntas de pesquisa e, de outro lado, os trabalhos que utilizam a Linguística de Corpus como suporte metodológico, ou como ferramenta para levantar dados (a serem analisados com base em outras teorias linguísticas).

As seções a seguir apresentam algumas sugestões para a definição de títulos e palavras-chave, de forma a aumentar a visibilidade da Linguística de Corpus em trabalhos acadêmicos, diferenciando as referências a corpora eletrônicos (da forma como a Linguística de Corpus trabalha) com outros tipos de corpora (como corpora literários não eletrônicos, não compilados a partir de certos critérios e não necessariamente analisados por meio de recursos da Linguística de Corpus, assim como corpora de pesquisas acadêmicas em geral, não necessariamente linguísticas).

5.1 Para trabalhos que investigam a Linguística de Corpus

Como anteriormente mencionado, optamos por trabalhar com duas vertentes: nesta primeira, fazemos referência aos trabalhos que têm a Linguística de Corpus como ponto central de suas investigações. Em casos assim, sugerimos:

a) Evitar o uso de siglas nos títulos e palavras-chave (uma vez que siglas muitas vezes funcionam apenas dentro da organização interna do trabalho) e a adoção do termo 'Linguística de Corpus';

b) Evitar, também, simplificações que utilizam apenas 'corpus' ou 'corpora' – considerando que trabalhos acadêmicos em geral (e não apenas os em linguística ou em Linguística de Corpus) têm um objeto de pesquisa que pode ser denominado corpus de estudo. Utilizar apenas 'corpus' dificulta a diferenciação da área em relação às demais áreas acadêmicas;

c) Fazer menção clara, no título do trabalho, à Linguística de Corpus – que funcionaria como uma área guarda-chuva, abrangendo tópicos de estudo relacionados;

d) Adotar o termo 'Linguística de Corpus' como primeira palavra-chave, reservando a segunda palavra-chave para apontar especificidades da investigação (como "colocações"; "prosódia semântica", etc.) e a terceira palavra-chave para esclarecer a natureza do corpus (paralelo, comparável, bilíngue, monolíngue) e a língua (ou par linguístico) . Em havendo a possibilidade de informar mais palavras-chave, utilizá-las para especificar o(s) texto(s) ou o(s) gênero(s) textual(is) que compõe(m) o corpus investigado (como *Dubliners*, ou corpora de textos jornalísticos etc.). O DIAGRAMA 1, a seguir, representa graficamente essa sugestão de hierarquização de palavras-chave para os trabalhos acadêmicos que investigam temas relacionados à Linguística de Corpus propriamente dita.

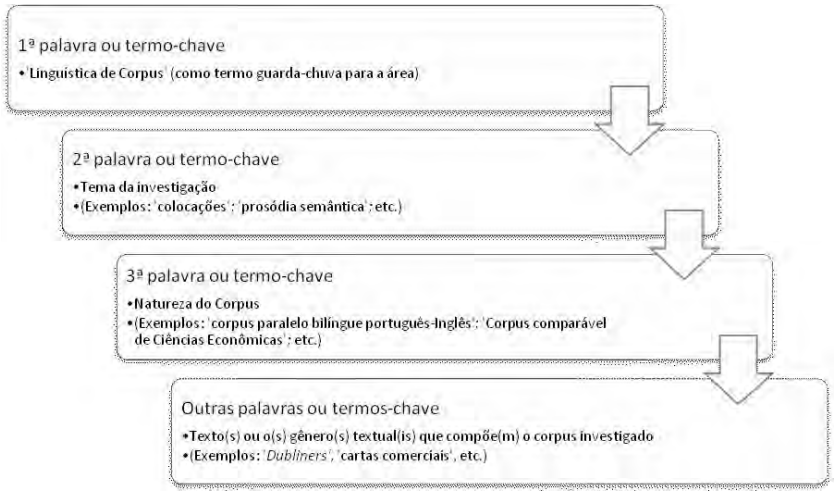


DIAGRAMA 1
Sugestão de hierarquização de palavras-chave para trabalhos que têm a Linguística de Corpus no cerne de suas investigações

5.2 Para trabalhos que utilizam a Linguística de Corpus como suporte metodológico

Nesta segunda vertente de trabalhos acadêmicos relacionados à Linguística de Corpus – utilizando-a como suporte metodológico para o levantamento de dados (a serem analisados por meio de outras teorias linguísticas) –, sugerimos:

a) Evitar o uso de siglas e o uso de simplificações (apenas 'corpus' ou 'corpora'), e adotar o termo 'Linguística de Corpus' para identificar a área, nos títulos ou nas palavras-chave;

b) Se cabível, fazer menção clara, no título do trabalho, à Linguística de Corpus;

c) Adotar o termo 'Linguística de Corpus' como segunda palavra-chave, reservando a primeira para a área em que o trabalho se insere (como, por exemplo, Tradução, Ensino, Terminologia, Descrição da Linguagem, etc.) e a terceira palavra-chave para apontar especificidades da investigação (como 'glossário de termos jurídicos', 'ensino de colocações', etc). Em havendo a possibilidade de informar mais palavras-

chave, utilizá-las para apresentar informações sobre a língua trabalhada e especificidades do corpus investigado. O DIAGRAMA 2, a seguir, representa graficamente essa sugestão de hierarquização de palavras-chave para os trabalhos acadêmicos que utilizam a Linguística de Corpus como ferramenta metodológica para o levantamento de dados (a serem analisados por meio de outras teorias linguísticas).

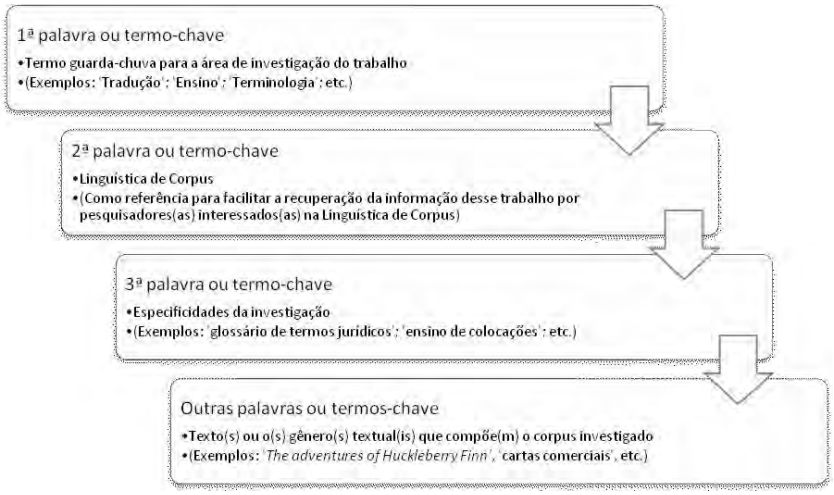


DIAGRAMA 2
Sugestão de hierarquização de palavras-chave para trabalhos que utilizam a Linguística de Corpus como ferramenta metodológica

6 Considerações finais

Este trabalho foi desenvolvido a partir da preocupação com a crescente dificuldade em se recuperar informações – dado o atual contexto mundial de grande produção de dados –, especialmente no que diz respeito a informações acadêmicas sobre a Linguística de Corpus.

Inicialmente, foi desenvolvida a análise de uma amostra de 84 trabalhos acadêmicos que têm suporte teórico-metodológico da Linguística de Corpus, visando identificar se, nesses trabalhos, eram claras as informações que associassem esses textos à Linguística de Corpus. Na análise, no entanto, foram identificados casos de trabalhos acadêmicos que não apresentam, claramente, suas afiliações à Linguística de Corpus.

Considerando que a falta de associação clara pode ter impacto direto sobre a recuperação desses trabalhos por estudiosos(as) interessados na área, foi sugerida a adoção de um vocabulário padrão e de uma metodologia para definir títulos e palavras-chave de trabalhos acadêmicos que se apoiam na Linguística de Corpus – tendo-a no cerne de suas pesquisas ou apenas utilizando-a como ferramenta metodológica para o levantamento de dados.

A análise aqui desenvolvida também reiterou o poder da Linguística de Corpus como ferramenta teórico-metodológica para a pesquisa linguística, tendo sido identificadas cinco grandes áreas que fazem uso da Linguística de Corpus, a saber: a) Ensino; b) Tradução; c) Terminologia; d) Descrição da linguagem; e e) Processamento de Linguagem Natural (PLN).

Também se destacaram, na análise aqui realizada, o interesse pela investigação acadêmica da língua inglesa – o que pode ser visto como um reflexo do fato de essa língua ter, na atualidade, o status de língua franca – e a importância dos textos literários para a formação de corpora investigados na área.

Esperamos, com a amostra dos estudos acadêmicos na área da Linguística de Corpus aqui analisada, ter ao menos trazido à baila o problema da recuperação da informação sobre a Linguística de Corpus nesses trabalhos. Nossas sugestões, no sentido de um vocabulário-padrão e de uma hierarquização das palavras-chave visa trazer maior visibilidade à nossa área, fazendo jus a sua relevância, tanto como objeto de pesquisa em si, quanto como abordagem em várias outras áreas, como o Ensino, a Tradução e a Terminologia, para citar apenas as mais evidentes.

Referências

BRAGA, Luis Paulo Vieira. *Introdução à mineração de dados*. 2. ed. Rio de Janeiro: E-Papers Serviços Editoriais, 2005.

CAMPOS, Maria Luiza de Almeida; GOMES, Hagar Espanha. Taxonomia e classificação: o princípio de categorização. *Revista de Ciência da Informação*, v. 9, n. 4, agosto de 2008. Disponível em: http://goo.gl/FmqCU. Acesso em: 4 de junho de 2012.

HILBERT, Martin; LÓPEZ, Priscila. The World's Technological Capacity to Store, Communicate, and Compute Information. *SCIENCE*, v. 332, 1º de abril de 2011.

II LINGUÍSTICA DE CORPUS E SEMÂNTICA

Pensar a modalidade: marcação epistêmica e evidencial no português brasileiro falado

Luciana Beatriz Avila[1]

RESUMO: A modalidade pode ser tomada como a avaliação de um conceptualizador sobre o material locutivo enunciado. Entretanto, definir essa categoria e seu *locus* de aplicação é uma tarefa árdua, dado que ela pode ser considerada uma categoria semântica se se assume que seu escopo é a proposição (cf. BYBEE; FLEISCHMAN, 1995), ou uma entidade pragmática, sendo o enunciado, nesse caso, seu domínio de aplicação (cf. MORAES, 1995). Muito dessa controvérsia, no nosso ponto de vista, se deve à teoria da linguagem ou à diamesia adotada pelos teóricos. Para fins deste trabalho, discutimos a modalidade a partir da perspectiva do uso real da língua, assumindo que ela é, na verdade, uma categoria semântica que pode ser usada de diferentes formas e pode, portanto, passar por diversos processos linguísticos, inclusive a gramaticalização. Para enfrentar esta miríade de usos, nosso trabalho tem como objetivo analisar os verbos de caráter epistêmico, ou, como vários autores descrevem (URMSON, 1969; HALL, 1958; HOOPER, 1975; HÜBLER, 1983; VENIER, 1991; SCHNEIDER, 1999), verbos de atitude proposicional, em um corpus de fala espontânea do Português Brasileiro.

PALAVRAS-CHAVE: modalidade epistêmica, verbos de atitude proposicional, corpus de fala espontânea, Português Brasileiro.

ABSTRACT: Modality in speech can be taken to be a conceptualizer's evaluation of an uttered locutive material. However, defining this category and its *locus* of application is a difficult task, given that it can be taken as a semantic entity if it is assumed that its scope is the proposition (cf. BYBEE; FLEISCHMAN, 1995), or a pragmatic

[1] Doutoranda no PosLin/UFMG, sob orientação da Profª Heliana Mello, com foco de pesquisa em estudos linguísticos baseados em corpora. Professora Assistente do Departamento de Letras da Universidade Federal de Viçosa, na área de Língua Portuguesa. Integra o grupo de pesquisa InCognito. E-mail para contato: luciana.avila@ufv.br.

entity, the utterance being, in this case, its domain of application (cf. MORAES, 1993). Much of this controversy, in our point of view, is due to the theory of language or the diamesy adopted by theoreticians. For the purpose of this paper, we discuss modality from the perspective of actual language use, assuming that it is in fact a semantic category that can be used in different ways, and can, therefore, undergo different linguistic processes, including grammaticalization. In order to tackle with this myriad of uses, our work aims to analyze epistemic verbs or, as many authors describe (URMSON, 1969; HALL, 1958; HOOPER, 1975; HÜBLER, 1983; VENIER, 1991; SCHNEIDER, 1999), verbs of propositional attitude, of a Brazilian Portuguese (BP) spontaneous speech corpus.

KEYWORDS: epistemic modality, verbs of propositional attitude, spontaneous speech corpus, Brazilian Portuguese.

1 Introdução

A noção de modalidade vem sendo debatida, ao longo da história, por filósofos, lógicos, linguistas. Diferentes autores tratam o tema de maneira diversa, fundados, principalmente, na ideia comum de que a modalidade está comprometida com a noção de verdade e com a opinião do falante sobre o que enuncia uma determinada proposição. No entanto, a tarefa de definir o que é essa categoria é difícil, porque, por exemplo, é um conceito ainda controverso (e complexo), que se pode sobrepor a outros como os de atitude e ilocução (cf. MELLO; RASO, 2012).

Alinhando-me com a proposta de Cresti (2002) e Tucci (2007) e seguindo a tradição balliniana, tomo a **modalidade** como **a avaliação por parte de um sujeito conceptualizador**[2] **do material locutório enunciado, ancorado em uma situação comunicativa**. Não estamos lidando mais com a categoria da modalidade em textos escritos, mas na fala, assim, não estamos mais no escopo da sintaxe, mas da

[2] O conceptualizador corresponde, em primeiro lugar, ao falante; em segundo, ao ouvinte/endereçado e, derivativamente, a uma terceira pessoa cuja perspectiva é levada em conta.

pragmática, o que significa que a modalidade vai incidir sobre uma unidade informativa/enunciado,[3] não uma proposição.

Para começar a enfrentar tal complexidade, este trabalho tem como objetivo analisar os verbos de caráter epistêmico, ou, como vários autores descrevem (URMSON, 1969; HALL, 1958; HOOPER, 1975; HÜBLER, 1983; VENIER, 1991; SCHNEIDER, 1999), verbos de atitude proposicional, em um corpus de fala espontânea do Português Brasileiro.

Na próxima seção, levanto o problema da sobreposição dos níveis semântico e pragmático, o que reflete nas análises que vêm sendo empreendidas, e também aponto para a fronteira entre modalidade epistêmica e evidencialidade, que será detalhada na seção 3.2.

2 Verbos de atitude proposicional e corpus

Mello, Carvalho e Côrtes (2010, p. 4) reconhecem que um dos problemas constantemente levantados nos estudos sobre a modalidade é o seu lugar de aplicação: se seria uma entidade semântica, uma vez que a categoria modalizável seria a proposição (cf. BYBEE; FLEISCHMAN, 1995), ou se seria uma entidade pragmática, sendo o enunciado, nesse caso, o seu campo de aplicação (cf. MORAES, 1993). Consideramos que muito dessa controvérsia passa pela definição mais geral de modalidade que confere ao falante/emissor/locutor a "responsabilidade", digamos, pela atribuição de valor a uma proposição, a princípio, neutra. Todas as diferentes acepções de modalidade mais correntes nascem no seio da semântica formal, que se compromete, como sabemos, com o julgamento das condições de verdade de sentenças declarativas. Ora, se o que se pretende aqui é discutir a modalidade no domínio discursivo, a partir do uso efetivo da língua, será necessário pensar não só como se dá o deslocamento da proposição para o enunciado,[4] mas também pensar

[3] Na moldura da Teoria da Língua em Ato (CRESTI, 2000), o enunciado é a unidade de referência da língua falada e é definido como qualquer expressão linguística interpretável pragmaticamente, ligada a: (a) uma condição semântica de plena significação da expressão em questão e (b) uma realização entoada segundo um padrão melódico de valor ilocutório.

[4] Segundo Cresti (2000), a proposição está para a escrita e para a sintaxe, assim como o enunciado está para a fala e para a pragmática.

sobre as categorias que pertencem, a rigor, a dimensões diferentes como os níveis pragmático e sintático-semântico que, em muitos estudos, se sobrepõem, principalmente as de falante / sujeito / agente.

Para ilustrar essa miríade de usos, tomo como exemplo os verbos de caráter epistêmico ou verbos de crença. Eles funcionam, segundo Venier (1991, p. 68 *apud* TUCCI, 2007, p. 172), como "sinais, para manifestar no ouvinte o grau de confiabilidade conferido pelo falante à proposição e para manter uma função sinalizadora também quando o enunciado que a contém vem reportado".[5] Vejamos:[6]

(i) os direitos com facilidade ou tem ciúme? João Ubaldo – Realmente não gostei de O Sorriso do Lagarto, mas gostei da adaptação de Sargento Getúlio para o cinema. Não tenho ciúme algum. Trata-se mesmo de outra obra, a obra do cineasta. iBEsp_242## **15 de outubro de 1997 Nahas se considera "um bode expiatório"** Estado – O sr. esperava esta sentença da Justiça? Naji Nahas – Não esperava de jeito nenhum. É mais uma violência, uma discriminação odiosa, entre as muitas que tenho sido vítima desde 89. Estado – Se o sr. é a vítima, quem (CDP:19Or:Br:Intrv:ISP)

(ii) Nahas – Confio na Justiça e sei que esta sentença está sujeita a revisão. Terminando meu trabalho, eu volto porque quero me apresentar para a apelaçao, aproveitando o momento para esclarecer todo esse assunto e mostrar definitivamente quem são os culpados. **Sou a vítima. Estado – O sr. Se considera um bode expiatório? Nahas – Sou. E o mais**

[5] Tradução minha para "segnali, di manifestare all'ascoltatore il *grado di attendibilità assegnato dal parlante* alla proposizione e di mantenete uma funzione segnaletica anche qando l'enunciato che li contiene viene riportato" (VENIER, 1991, p. 68 *apud* TUCCI, 2007, p. 172).

[6] Os exemplos foram coletados do Corpus do Português, esse é um corpus de referência da língua portuguesa, projeto desenvolvido pelos professores Mark Davies (BYU) e Michael J. Ferreira (Georgetown University), com mais 45.000.000 de palavras, e um total de 57000 textos do séc. XIV ao séc. XX. Esse corpus permite o cruzamento de dados e distribuição de palavras, frases e construções por **registro** (oral, ficção, jornalístico e acadêmico); **dialeto** (português brasileiro vs europeu no século XX); **período histórico** (séculos XIV ao XX). Para acessá-lo: http://www.corpusdoportugues.org.

> **sofrido do Brasil**.iBEsp_244## 18 de outubro de 1997 Manoel de Barros faz do absurdo sensatez Estado – Como surgiu seu amor pelas coisas sem importância? Manoel de Barros – Quando eu era jovem, fiz uma longa viagem pela Bolívia. (CDP:19Or:Br:Intrv:ISP)

Em (i), confirmado por (ii), tem-se um caso de discurso reportado e mostra-se a separação entre "quem enuncia" e, nos termos tradicionais explicitados anteriormente, "o falante que se compromete com a verdade da proposição enunciada". Em última consequência, há uma separação, a meu ver, entre enunciado e proposição enunciada. Observemos outro exemplo:

(iii) lugar (exceto Belo Horizonte) por mais de 2 anos. Sebastião: Nem no Rio de Janeiro? Prof. Eduardo: No Rio, eu peguei uma vez, mas minha mulher me gozou tanto. **Comecei a puxar o "s", igual ao pessoal de Juiz de Fora, que se considera carioca.** Dizem aqui em Belo Horizonte que o pessoal de lá dá o endereço tipo Avenida Brasil, 9 milhões, 582 mil etc. (risos) Foi na época em que servi o Exército no Rio em Magalhães Bastos, subúrbio. Sebastião: Não entendi. Com 17 anos o senhor

A partir de (iii), temos duas possíveis interpretações:

(iiia) Os juizforanos efetivamente dizem que são cariocas (e aí se constituiria, como (xi), discurso reportado)

(iiib) Os juizforanos se comportaram de uma determinada forma ou fizeram alguma consideração sobre serem cariocas que levam o falante a crer que o pessoal de Juiz de Fora "se considera carioca" (inclusive enfatizado pelo enunciado seguinte em que temos a presença do evidencial "*Dizem que*").

Dessa forma, o que se pode depreender dos exemplos acima, quando empregada a terceira pessoa do discurso, é que um dos usos dessas construções é uma avaliação do falante sobre a avaliação (ou perspectiva ou, ainda, ponto de vista) de uma pessoa outra (no papel sintático de sujeito da cláusula principal). Também entendo que as noções de modalidade epistêmica e evidencialidade se confundem

nessas construções e que, talvez, não se constituam como categorias estanques, mas componham um *continuum* que parte de expressões de autoavaliação, passando pelo discurso relatado, em que o comprometimento do falante é mais opaco.

A seguir, detalho a metodologia empregada e, na seção 3, a partir de uma abordagem pragmático-cognitiva, analiso esses verbos de atitude proposicional em um corpus oral representativo do Português Brasileiro, em que tento lançar luz sobre as nuances acima apresentadas e apontar algumas tendências de comportamento dessas construções.

3 Metodologia

Metodologicamente, utilizamos os dados do C-ORAL-BRASIL (RASO; MELLO, 2010, 2012), um corpus de fala espontânea do PB, quinto braço do C-ORAL-ROM (CRESTI; MONEGLIA, 2005), um conjunto de corpora comparáveis representativo das quatro principais línguas românicas europeias (Italiano, Espanhol, Português e Francês), prosodicamente segmentado em enunciados e unidades tonais, de acordo com a Teoria da Língua em Ato (CRESTI, 2000), e alinhado pelo software WinPitch (MARTIN, 2000) – que permite o exame simultâneo de som, espectrograma e texto.

O C-ORAL-BRASIL representa a diatopia do estado de Minas Gerais, e fornece uma descrição diastrática balanceada. Tomamos como amostra um subcorpus composto por 20 textos de três tipologias interacionais, divididos em privados e públicos: 7 monólogos (5 privados e 2 públicos), 7 diálogos (5 privados e 2 públicos) e 6 conversações (4 privadas e 2 públicas), em um total de aproximadamente 25.000 palavras.

Foram encontradas 130 ocorrências de verbos de atitude proposicional (excluídos os verbos 'saber' e os verbos com perguntas encaixadas, tais como 'perguntar', 'falar', 'ver', 'olhar', 'contar', os três últimos na acepção de 'verificar'), em um total de 1152 enunciados modalizados, correspondente a 11,2% de todos os enunciados. Entre os verbos – o marcador modal mais frequente (55,2%) – esta estratégia corresponde a 20,3% das ocorrências.

4 Resultados e análise dos dados

Os types e tokens correlatos foram classificados quantitativamente de acordo com a tipologia interacional (público versus privado;

monólogos versus diálogos versus conversações), tipologia textual (narrativo, relato, expositivo, argumentativo, descritivo)[7] e o nível de escolaridade do falante (aproximadamente, nível 1: escola primária, nível 2: escola secundária, nível 3: universitário e educação profissionalizante).

Além disso, foi utilizado um número de variáveis na tabulação dos dados, que leva em conta: o padrão de estrutura informacional (qual unidade informacional[8] contém o marcador modal), composicionalidade (enunciado simples, dois índices no mesmo enunciado, dois ou mais

[7] A tipologia textual não é um dos parâmetros controlados no C-ORAL-BRASIL; portanto, não pode ser balanceado. Neste trabalho foi levado em consideração, a fim de se checar se essa variável afetava os resultados encontrados. Esta classificação é baseada em Dolz & Schneuwly (2004).

[8] A tabela abaixo descreve as unidades informacionais e suas funções (adaptada de Tucci, 2007):

	UI	Função informativa	Etiqueta
Unidades textuais	Comentário	Exprime a força ilocucionária do enunciado	COM
	Tópico	Especifica o campo de aplicação da força ilocucionária do comentário	TOP
	Parentético	Expressa uma integração metalinguística do enunciado, modalizadora, apresentando um ponto de vista externo àquele do comentário	PAR
	Introdutor locutivo	Sinaliza a suspensão pragmática do *hic et nunc* e introduz uma metailocução	INT
	Apêndice	Integra a unidade de comentário (ou tópico) com informação não-essencial	APC
Unidades dialógicas	Incipitário	Assinala a tomada de turno do falante	INP
	Fático	Regula o canal comunicativo	PHA
	Alocutivo	Especifica a quem a mensagem é dirigida e mantém a atenção do interlocutor	ALL
	Conativo	Incita o interlocutor a tomar parte da troca comunicativa	CNT
	Expressivo	Estimula o interlocutor a compartilhar um ponto de vista comum sobre o enunciado	EXP
	Conectores dialógicos	Assinala para o ouvinte que o turno terá continuidade	DCT

índices em enunciados diferentes, referência contextual[9]), tipo de modalidade e seus subvalores, o padrão sintático, a projeção pragmática dos índices, e o holder.[10]

Os tipos encontrados foram: 'achar', 'acreditar', 'crer', 'imaginar', 'julgar', como exemplificado abaixo:

(1) **Achar:**

(a) *GIL: [2] <ô /=CNT= mas> /=DCT= voltando à questão /=COB= falando em [/2]=EMP= e também falando em povo mascarado /=COB= esse povo do Galáticos é muito palha /=COB= eu acho que es nu deviam mais participar /=COM= e <tal> //=UNC=$ (*bfamcv01*)

(b) [44] aí /=PHA= passou um pouquim /=COB= o filho /=i-COB= achando que tava errado aquele negócio /=PAR= voltou lá outra vez //=COM=$ (*bfammn03*)

(c) *KAT: [43] então ela acha que é a meia que tá melhorando //=COM=$ (*bfamdl04*)

[9] A modalidade, de acordo com Tucci (2007), não é uma propriedade do enunciado, mas da unidade informacional. Há três casos a serem levados em conta: (a) se o enunciado é simples, istio é, se contem apenas a unidade de Comentário, a modalidade pertence simultaneamente a todo o enunciado e à unidade informacional; (b) se dentro de uma mesma unidade informacional temos dois índices modais, vale o *princípio de composicionalidade*, em que um índice domina o outro e determina a modalidade; e (c) se um enunciado possui duas unidades informacionais, se aplica um *princípio de não-composicionalidade*, e o domínio da modalidade, portanto, recai sobre a unidade informativa. Em nosso ponto de vista, consideramos que a modalidade não só incide localmente, no âmbito da unidade informacional, mas também é uma propriedade dinâmica, que atua igualmente, em uma cadeia anafórica, no domínio interacional.

[10] O holder expressa o conceptualizador que avalia um material locutório enunciado, que pode coincidir com o falante, o endereçado ou um outro indivíduo cuja perspectiva está em consideração.

(2) **Acreditar:**

(d) [46] tipo /=INT= eu [/1]=EMP= eu acredito /=i-COM= tipo /=PAR= cem /=SCA= por cento /=SCA= nisso //=COM=$ (*bfamcv01*)

(e) [87] não /=INP_r= nu acredito nisso não //=COM_r=$ (*bfamdl01*)

(f) [49] e eu acredito que depois que eu terminar o EDUCONLE /=COB= eu acho que aí eu vou tar mais madura ainda /=COB= acho que mais preparada //=COM=$ (*bpubmn01*)

(3) **Crer:**

(g) *ENC: [208] eu creio que sim //=COM (*bfamdl05*)

(4) **Imaginar:**

(h) [171] não /=PHA= trinta reais /=TOP= aí eu &j [/2]=SCA= eu [/1]=EMP= eu fico imaginando que e' fica pensando assim /=INT= Nossa Sio' /=EXP_r= às vezes lá em casa tá precisando de fazer uma compra e tudo /=COM_r= né //=PHA=$ (*bpubmn01*)

(5) **Julgar:**

(i) BAL: [169] existem vários> /=COB= só que a maioria /=TOP= &he /=EMP= tá julgando improcedência /=COB= tal //=COM=$ [170] porque /=TMP= &he /=EMP= de certa forma /=PAR= a bancada evangélica /=TOP= eles tão /=SCA= muito contra /=COM= essa coisa /=APC= né //=PHA=$ (*bfamdl02*)

Como mencionado anteriormente, o número total de tokens de verbos de atitude proposicional encontrado no subcorpus foi 130, distribuídos em 5 types, como a seguir: 123 para 'achar' (93,18%); 3 para 'acreditar'; 1 para 'crer'; 1 para 'imaginar'; 2 para 'julgar'. A Figura 1 mostra esta distribuição:

FIGURA 1
Distribuição dos verbos de atitude proposicional na amostra

Analisando os números para tipologia interacional, os resultados são: 31 exemplares para monólogos, 16 em monólogos públicos e 15 em privados; 63 exemplares para diálogos, 10 para públicos versus 53 para diálogos privados; e 36 exemplares para conversações, 6 em públicas versus 30 em conversações privadas. Podemos visualizar os dados na Tabela 1:

TABELA 1
Distribuição de tokens de verbos de atitude proposicional
por tipologia interacional

	Privado	**Público**	**TOTAL**
Monólogos	15	16	31
Diálogos	53	10	63
Conversações	30	6	36
TOTAL	98 (75,3%)	32 (24,7%)	130

Esta tabela aponta para uma alta taxa de tokens em interações dialógicas, principalmente em diálogos privados. Isso se deve a características específicas desses tipos de texto e à relação dos

participantes na situação comunicativa. Se levarmos em conta a tipologia textual, nota-se que 39,2% destes verbos são usados em textos argumentativos ou expositivos, o que coincide com a sua própria definição: sinalizar o comprometimento do conceptualizador em relação ao que se enuncia. As outras ocorrências distribuídas em textos narrativos, descritivos e relatos mostram os momentos em que os participantes expressam suas opiniões ou crenças ou requerem a opinião de seu interlocutor. Devo destacar que as ocorrências em monólogos públicos correspondem a apenas um texto: a participante, uma professora da escola básica relata sai atividade profissional e dá sua opinião sobre o processo de ensino-aprendizagem. É também necessário destacar que a diferença entre os números de textos privados e públicos é minimizada em termos de frequência relativa.

4.1 Padrões sintáticos, semântica e questões pragmáticas

Vários padrões sintáticos são utilizados, principalmente:

(i) os verbos de atitude proposicional introduzem orações encaixadas. Em nossa amostra, 76,15% de todas as ocorrências seguem este padrão.

(ii) eles podem ocupar diferentes posições no enunciado, nos casos em que não introduzem uma encaixada.

(iii) $[SN]_{SUBJ}$ V $[SN]_{OBJ}$ [SAdv] $[SAdj]_{ATR.}$

Alguns exemplos:

(j) SN_{SUJ} V comp S
*LUI: [236] eu acho que a gente deve chamar os <times> legais //=COM=$ (*bfamcv01*)

(k) SN_{SUJ} V SAdv S
*SIL: [154] eu acho assim /=INT= se a pessoa nu tem condições de fazer /=TOP= ele paga pra fazer //=COM=$ (*bfamdl04*)

(l) SN_{SUJ} V SN_{OBJ} SAdv $SAdj_{ATR}$
[77] então eu achava aquilo muito interessante //=COM=$ (*bfammn06*)

(m) SN_{SUJ} $SAdv_{NEG}$ V $SAdj_{ATR}$ $SAdv_{NEG}$
[283] eu nu achei ruim não /=COM= Jael //=ALL=$ (*bfamcv02*)

Semanticamente, estes verbos expressam o grau de comprometimento do conceptualizador em relação ao material locutivo enunciado. Estas unidades lexicais estão ligadas a diferentes frames descritos para o inglês: Opinion, Certainty, Awareness, Assessing and/or Cogitation.[11]

Em termos da distribuição dos verbos por unidades informacionais, podemos ver na Figura 2 como o subcorpus está caracterizado:

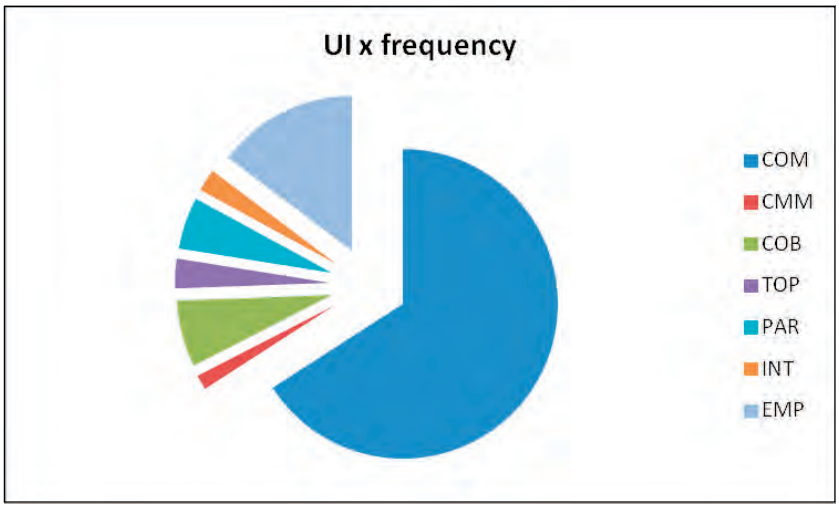


FIGURA 2
Distribuição dos verbos de atitude proposicional
em unidades informativas

[11] A descrição para o frame de Opinion, por exemplo, é a seguinte: "Cognizer holds a particular Opinion, which may be portrayed as being about a particular Topic (John THINKS that it looks better back)". Disponível em: https://framenet.icsi.berkeley.edu/fndrupal/

A primeira coisa a observar é qual a unidade informacional que contem o índice modal: apenas o Comentário (incluindo as unidades Comentário Múltiplo e Comentário Ligado), o Tópico, o Parentético e o Introdutor Locutivo podem ser modalizados. No caso dos verbos de atitude proposicional, a unidade de Comentário é modalizada em 65,38% de todas as ocorrências, seguida pelo Parentético e o Tópico em um número bem baixo, 7 e 4, respectivamente. Confiram os exemplos:

(o) **COM**

[94] **achei aquele lugar incrível** //=COM=$ (*bfamcv01*)

(p) **TOP**

*ANE: [324] **eu acho que quando eu vim** /=TOP= tinha esse /=CMM= tinha o outro //=CMM=$ (*bfamdl05*)

(q) **PAR**

[41] só que é de microondas /=COM= **eu acho** //=PAR=$ (*bfamdl01*)

(r) **INT**

*SIL: [154] **eu acho assim** /=INT= se a pessoa nu tem condições de fazer /=TOP= ele paga pra fazer //=COM=$ (*bfamdl04*)

Mello e Raso (2012, a aparecer) discutem a necessidade de ampliação do enunciado e das unidades tonais como construtos analíticos básicos no tratamento da fala espontânea via frames. Como, segundo a Teoria da Língua em Ato (CRESTI, 2000), ao enunciado corresponde uma ilocução, mesmo que esses verbos epistêmicos, objeto de nossa investigação, estejam, de alguma forma, representados em frames ligados por herança metafórica e no uso de atividade mental, ao falarmos, estamos cumprindo diferentes ações. Dessa forma, "a análise da linguagem pautada por frames, necessariamente deve levar em conta pelo menos dois níveis adicionais àqueles construcional e semântico, quais sejam, os níveis informacional e ilocucionário (MELLO; RASO, 2012, p. 12).

Transponho um princípio talhado, até agora, para a análise da escrita para a análise da fala, o Princípio da Não-Sinonímia, para discutir brevemente essa questão. Segundo esse Princípio (GOLDBERG, 1995, 2006):

> Se duas construções são sintaticamente distintas, elas devem ser semântica ou pragmaticamente distintas.
>
> *Corolário A*: Se duas construções são sintaticamente distintas e semanticamente sinônimas, então elas não devem ser pragmaticamente sinônimas.
>
> *Corolário B*: Se duas construções são sintaticamente distintas e pragmaticamente sinônimas, então elas não devem ser semanticamente sinônimas.

Os verbos, próximos semanticamente, se organizam sintaticamente de várias maneiras, como vimos. Dessa forma, pelo PNS, o seu comportamento pragmático deve ser diferente. E de fato o é. Vejamos as possíveis funções e os seus respectivos exemplos com a mesma unidade lexical 'achar':

(i) Estrutura informacional: em posição parentética funciona como atenuador da asserção anterior:

Contexto: casal conversa sobre a distribuição de vagas em uma universidade pública.

*LAU: [51] não / tem um de ensino de arte +
*LUZ: [52] são duas vagas / eu acho // [53] não// [54] de ensino de artes // (*bfamdl03*)

(ii) Estratégia de polidez positiva: mitigação de assimetria social.

Contexto: mulher de classe média quer comprar um apartamento. Ela conversa com um dos operários na obra.

***ANE: [207] <só tem esse> apartamento pra vender //**
*CES: [208] <obrigado> //
***ENC: [209] eu creio que sim // [210] que +**
*ANE: [211] tá //
*ENC: [212] tá // (*bfamdl05*)

Apesar de não termos uma medida de comparação sintático-semântica para o verbo 'crer' no subcorpus, uma outra variável considerada, a relação entre os participantes, nos dá o caminho para a projeção pragmática desse verbo epistêmico de crença: ele pode ser usado como uma estratégia de polidez, na tentativa de minimizar a relação hierarquicamente assimétrica entre os interactantes. Na perspectiva de Brown and Levinson (1987), essa seria uma estratégia de polidez positiva, em que o falante reivindica uma base comum com o interlocutor, no caso, a utilização do que considera um marcador de identidade de grupo.

(iii) Marcadores de concordância / discordância: indica um (possível) padrão lexical. Corresponde a um fenômeno de gradiência de índice modal para um marcador discursivo:

Contexto: amigas no supermercado.

*REN: [413] <pode> // [414] tanto faz // [415] pode //
*FLA: [416] ou cê acha muito //
*REN: [417] uhn // [418] acho que não // [419] tá //
[420] papel <higiÃªnico / eu nunca> + (*bfamdl01*)

Dessa forma, damos conta de que uma descrição de frames em termos exclusivamente sintáticos e semânticos pode ser, de fato, enriquecida com desdobramentos pragmáticos de diversas ordens.

4.2 Modalidade epistêmica x evidencialidade

A relação semântica entre modalidade epistêmica e evidencialidade ocupa o debate entre linguistas (CHAFE, 1986; FITNEVA, 2001; NUYTS, 2001; PLUNGIAN, 2001, entre outros) e nem sempre está clara na literatura a fronteira entre as duas noções. Dendale e Tasmowski (2001, p. 341-342) levantam os problemas de conceituação e apontam três caminhos encontrados em estudos recentes: "*disjunção* (em que são conceptualmente distinguidas), *inclusão* (em que uma está incluída no escopo semântico da outra) e *sobreposição* (em que elas parcialmente se interseccionam).[12]

[12] Tradução minha para: "*disjunction* (where they are conceptually distinguished from each other), *inclusion* (where one is regarded as falling within the semantic scope of the other), and *overlap* (where they partly intersect)".

A modalidade epistêmica expressa os diferentes graus de comprometimento em relação à validade do material enunciado. Já os evidenciais são marcadores que indicam a fonte e a confiabilidade do conhecimento do falante. Apesar de a relação entre as duas ser bastante evidente, quando aplicada a dados de fala, se mostra empiricamente ainda mais insatisfatória. Como as análises são centradas no falante, alguns problemas emergem.

Em nossa amostra, levanto alguns exemplos que podem clarear a questão e sugerir que as categorias não estão em oposição, não estão em relação de inclusão, nem se encontram em um ponto de intersecção, mas se constituem em um *continuum*, a depender se está em foco um conceptualizador primário ou se se trata de um conceptualizador em terceira pessoa. Vou partir de três exemplos para evidenciar meu ponto.

(s) *LUI: [236] eu acho que a gente deve chamar os <times> legais //=COM=$ (*bfamcv01*)

(t) [65] se o brasileiro nu lê os manuais /=TOP= hhh no mercado de reposição /=TOP= &auto [/1]=SCA= de autopeça /=APT= eles acham que abrir uma empresa é comprar um produto por um real /=COB= na base cem /=COB= e vender por dois acha que tá ganhando o &do [/2]=SCA= o dobro //=COM=$ (*bfammn06*)

(u) *ROG: [187] dá // [188] tem muita pedra ali / uai // [189] lá embaixo ainda tem pedra / perto da [/2] perto da garagem / lá tem // [190] assim vai ficar uma pracinha boa aqui //
*PAU: [191] e a Isa tava achando que ela ali ia ficar pequena // [192] falei assim / depois de feito é que a gente vê o tamanho / né // (*bpubdl01*)

No exemplo em (s), temos um julgamento epistêmico de um conceptualizador primário (o falante), o mais frequente em nossa amostra. Em (t), há a ocorrência de uma avaliação por parte do falante de um julgamento epistêmico de um conceptualizador em terceira pessoa, o que poderia ser textualizado como "eu acho que eles acham". Por fim, em (u), o falante reporta um julgamento epistêmico de uma terceira pessoa, porque partilha dessa informação (ele/ela acha, porque compartilhamos esse conhecimento).

O seguinte *continuum* representa os graus de comprometimento/ afastamento do falante, conceptualizador primeiro, no que diz respeito ao material enunciado:

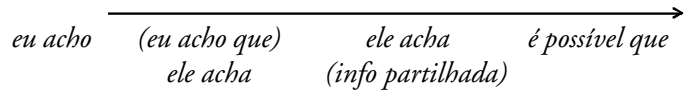


FIGURA 3
Continuum modalidade epistêmica/evidencialidade

Nos exemplos, observamos três diferentes graus de participação do enunciador na constituição da avaliação sobre o que está sendo dito, o que me leva a tentar reformular o conceito de modalidade e assumir a noção de conceptualizador, nos termos langackerianos, para explicar este fenômeno semântico.

5 Considerações finais

A partir de análises qualitativas e quantitativas, observou-se que o verbo epistêmico mais frequente é o 'achar', responsável por 93,18% de um total de 130 ocorrências.

Ainda, em termos sintáticos, um número de padrões é usado e o mais comum é o verbo como introdutor de uma oração encaixada. Semanticamente, estes verbos apontam para os diferentes graus comprometimento de um conceptualizador em relação ao material locutivo enunciado.

Deduzimos que, nos níveis pragmático e discursivo, os verbos podem: (a) sinalizar, em posição parentética, a atenuação de uma asserção anterior; (b) funcionar como marcadores de concordância e discordância, o que sugere um caminho de gramaticalização em índices de modalização; (c) servir como uma maneira de mitigar assimetrias socioculturais entre os participantes da interação. Os distintos comportamentos pragmáticos nos leva a refletir sobre a necessidade de se ampliar a anotação de frames para uma camada dessa natureza.

Finalmente, estes marcadores modais podem levantar a questão das fronteiras entre a modalidade epistêmica e a evidencialidade, com a proposta de um *continuum* para representar o nível de responsabilidade do conceptualizador primário sobre uma determinada avaliação ou sobre a fonte de informação.

Referências

AUSTIN, J. *How to do things with words*. London: Oxford University Press, 1962.

BROWN, P.; LEVINSON, S. *Politeness*: some universals in language usage. Cambridge: Cambridge University Press, 1987.

BYBEE, J.; FLEISCHMAN, S. *Modality and grammar in discourse*. Amsterdam / Philadelphia: John Benjamins, 1995.

CHAFE, W. Evidentiality in English conversation and academic writing. In: CHAFE, W.; NICHOLS, J (Ed.). *Evidentiality*: The Linguistic Coding of Epistemology. New York: Ablex, 1986, p. 261-272.

CRESTI, E. *Corpus di italiano parlato*. Firenze: Accademia della Crusca, 2000.

CRESTI, E. Illocuzione e modalità. In: BECCARIA, P.; MARELLO, C. (Ed.). *La parola al testo*. Scritti per Bice Mortara-Garavelli. Torino: Ed. dell'Orso, 2002, p. 133-145.

CRESTI, E.; MONEGLIA, M. *C-ORAL ROM*: Integrated reference corpora for spoken Romance languages. Amsterdam/Philadelphia: John Benjamins, 2005.

DAVIES, M.; FERREIRA, M.. (2006-). *Corpus do Português* (45 million words, 1300s-1900s). Disponível em: http://www.corpusdoportugues.org. Último acesso em: 30 mai. 2012.

DENDALE, P.; TASMOWSKY, L. Introduction: evidentiality and related notions. *Journal of Pragmatics*, 33, 339-348, 2001.

DOLZ, J.; SCHNEUWLY, B. *Gêneros orais e escritos na escola*. Campinas: Mercado de Letras, 2004.

FITNEVA, S. Epistemic marking and reliability judgments: evidence from Bulgarian. *Journal of Pragmatics*, 33, 401-420, 2001.

GOLDBERG, A. *Constructions*: a construction grammar approach to argument structure. Chicago: University of Chicago Press, 1995.

GOLDBERG, A. *Constructions at work*. Chicago: University of Chicago Press, 2006.

HALL, R. Assuming: on the set of positing words. *Philosophical Review*, 67, p. 52-75, 1958.

HOOPER, J. B. On assertive predicates. In: KIMBALL, J. P. (Ed.). *Syntax and semantics*, 4. New York / London: Academic Press, 1975, p. 91-124.

HÜBLER, A. *Understatements and hedges in English*. Amsterdam / Philadelphia: John Benjamins, 1983.

MARTIN, P. WinPitch Corpus. Disponível em: http://www.winpitch.com.

MELLO, H. R.; CARVALHO, J. M.; CORTES, P. O. Modalização na fala espontânea do português brasileiro: um primeiro mapeamento de índices morfolexicais. *Revista de Estudos da Linguagem*, v. 18, p. 105-133, 2010.

MELLO, H. R.; RASO, T. Illocution, modality, attitude: different names for different categories. In: MELLO, H. R.; PANUNZI, A.; RASO, T (Ed.). *Pragmatics and prosody*: illocution, modality, attitude, information patterning and speech annotation. Firenze: Firenze University Press, 2011, p. 1-18.

MELLO, H.; RASO, T. *Frames e fala espontânea*. 2012. (Manuscrito submetido à publicação).

MORAES, J. A. de. A entoação modal brasileira: fonética e fonologia. *Cadernos de Estudos Linguísticos*, 25, Campinas: UNICAMP, 1993.

NUYTS, J. *Epistemic modality, language and conceptualization:* a cognitive-pragmatic perspective. Amsterdam/Philadelphia: John Benjamins, 2001.

PLUNGIAN, V. The place of evidentiality within the universal grammatical space. *Journal of Pragmatics*, 33, 349-357, 2001.

RASO, T.; MELLO, H. The C-ORAL-BRASIL corpus. In: MONEGLIA, M.; PANUNZI, A. (Org.). *Bootstrapping information from corpora in a cross linguistic perspective*. Firenze: Firenze University Press, 2010, p. 193-213. Available at: http://www.fupress.com/Archivio/pdf%5C4106.pdf.

RASO, T.; MELLO, H. *C-ORAL-BRASIL I*: Corpus de referência do português brasileiro falado informal e DVD multimedia. Belo Horizonte: Editora UFMG, 2012. v. 1.

SCHNEIDER, S. *Il congiuntivo tra modality e subordinazione*. Roma: Carocci Editore, 1999.

TUCCI, I. *L'espressione lessicale della modalità nel parlato spontaneo*: Analisi del corpus C-ORAL-ROM. 2007. Tese (Doutorado – Linguística Italiana). Firenze, Università degli Studi di Firenze, 2007, 335p.

TUCCI, I. L'espressione lessicale della modalità nel parlato spontaneo: Analisi del corpus C-ORAL-ROM. 2007. Tese (Doutorado – Linguística Italiana). Firenze, Università degli Studi di Firenze, 2007, 335p.

URMSON, J. O. Parenthetical verbs. *Mind* 61: 480-496, 1969.

VENIER, F. *La modalizzazionne assertiva.Avverbi modalie verbi parentetici*. Milano: Franco Angeli, 1991.

Formas modais equivalentes com valores semânticos diversos: um mapeamento em corpora

Raíssa Caetano[1]
Luis Filipe Lima e Silva[2]

RESUMO: O estudo da modalidade apresenta certa tradição nos estudos linguísticos e, por essa razão, este fenômeno vem sendo estudado sob diferentes visões teóricas. O objetivo deste trabalho é propor um estudo da modalidade na fala espontânea, a partir de uma perspectiva de base empírica inovadora na literatura sobre o tema. Para tanto, observarmos a ocorrência de duas variantes de marcadores de modalidade, uma adverbial e outra predicativa, para discutir a equivalência semântica no que diz respeito ao valor modal, isto é, a possibilidade de uma dada sinonímia. Os itens investigados são os pares *realmente/na realidade, obviamente/é óbvio, claramente/é claro* e *logicamente/é lógico*. Os dados foram analisados a partir dos *corpora* C-ORAL-BRASIL (RASO & MELLO, 2012) e *Corpus* do Português (DAVIES & FERREIRA, 2006). Constatou-se que os pares não são semanticamente correlatos, de modo que não são concorrentes, mas coocorrentes na língua. Apresentaram uma variação condicionada sintática, semântica e pragmaticamente. O fator pragmático influencia no emprego dos modais, tema que deve ser mais bem investigado.

PALAVRAS-CHAVES: Modalidade. Marcadores Adverbiais. Sinonímia. Corpus. Semântica. Pragmática.

[1] Graduanda do 7º período do curso de Letras (Licenciatura dupla em Português-Italiano) da Universidade Federal de Minas Gerais. Bolsista do CNPq de Iniciação Científica sob orientação da professora Heliana Mello (UFMG). E-mail para contato: <raissavoliveira@gmail.com>

[2] Graduando do 7º período do curso de Letras (Bacharelado em Lingüística) da Universidade Federal de Minas Gerais. Bolsista da FAPEMIG de Iniciação Científica sob orientação da professora Heliana Mello (UFMG). E-mail para contato: <luis.1397@yahoo.com>

ABSTRACT: The study of modality has been focused in Linguistics for some time under different theoretical approaches. In this paper we study modality in spontaneous speech, supported by an innovative empirical perspective. In order to carry our study we looked at two variants of modal markers: an adverbial one and a predicative one, having as a goal the prospection of possible semantic equivalence, that is, the likelihood of there being synonymy between the two forms. The investigated pairs are: *realmente/na realidade, obviamente/é óbvio, claramente/é claro* and *logicamente/é lógico*. The data analysed were taken from the C-ORAL-BRASIL (RASO & MELLO, 2012) and the Corpus do Português (DAVIES & FERREIRA, 2006). Our analysis led to the conclusion that the pair are not semantic correlates, therefore they coocur in the language and are not in complementary distribution. Pragmatic constrains influence modal use, however this needs to be further investigated.

KEYWORDS: Modality. Adverbial Markers. Synonymy. Corpus. Semantics. Pragmatics.

1 Introdução

Neste trabalho, buscamos discutir a questão da sinonímia e como a análise de diversos níveis da língua a partir de dados sistematizados da língua oral auxilia na distinção entre diferentes usos dos marcadores de modalidade. Com o auxílio dos *corpora* orais[3] para a pesquisa linguística, a modalidade pode ser melhor compreendida, bem como ter seu escopo delimitado. Foi investigado o estatuto da equivalência entre duas variantes de marcadores de modalidade, a saber, "realmente/na realidade, obviamente/é óbvio, claramente/é claro e logicamente/é lógico". A presente pesquisa utilizou-se de dois *corpora*, o C-ORAL-BRASIL (RASO & MELLO, 2012) e o Corpus do Português (DAVIES & FERREIRA, 2006) como suporte para a investigação do referido fenômeno.

Os pares, constituídos por uma forma adverbial e uma predicativa, foram analisados quantitativa e qualitativamente. O advérbio *realmente* foi analisado diacronicamente, em uma

[3] Para um panorama da importância dos *corpora* para a pesquisa linguística, cf. Facchinetti (2007).

investigação a respeito de um possível processo de gramaticalização. Com relação aos demais pares, foram considerados aspectos semânticos, sintáticos e pragmáticos com o intuito de individualizar suas funções. O artigo foi organizado da seguinte maneira: na seção 1, apresentamos uma revisão da bibliografia sobre o tema, na seção 2 explicamos a metodologia empregada neste trabalho, na seção 3 apresentamos as análises dos dados e, por fim, na seção 4 mostramos as considerações finais.

2 Revisão bibliográfica

A definição de modalidade não está bem estabelecida na literatura sobre o tema. Há definições clássicas, como a de Bally (1932) que considera o *dictum* como o conteúdo da enunciação e o *modus* como a atitude do sujeito frente ao conteúdo proposicional. A maioria dos autores define modalidade utilizando o termo *atitude* (cf. HALLIDAY, 1970; SCHNEIDER, 1999; BYBEE, 1985; CRESTI, 2001). Mello & Raso (2012) problematizam esse tratamento a partir da discussão de três noções, são elas: atitude, ilocução e modalidade. O principal aspecto do estudo desenvolvido por esses autores a ser considerado nesta investigação é a importância da delimitação do nível linguístico no qual a modalidade atua. Mello (2009) afirma que modalidade é um

> domínio **semântico** que acomoda variadas nuances de sentido, adicionadas a uma hipotética estrutura neutra, qual seja, uma proposição factual e declarativa. Essa variedade de sentidos encobre um espectro de sub-classes que inclui conteúdos semânticos desiderativos, intentivos, hipotéticos, dubidativos, dentre outros (MELLO, 2009)[4]

Saber como a interface semântica-pragmática atua na expressão da modalidade ainda é uma tarefa árdua para os teóricos. Propomos que essa relação seja investigada a partir de dados do uso real da língua. Essa tarefa está sendo desenvolvida pelos pesquisadores que

[4] A definição encontra-se em uma proposta de desenvolvimento de projeto de pesquisa enviada ao CNPq em 1/2010, a qual se denomina *"Construções adverbiais modalizadoras na fala espontânea: um estudo prospectivo"*.

integram o projeto guarda-chuva *C-ORAL-BRASIL: formação de corpus e estudos sobre a fala espontânea do português do Brasil* (autorização COEP 0209.0.203.000-17), que, por sua vez, participam de um programa de pesquisa sobre o fenômeno da modalidade e sua expressão coordenado pela professora Heliana Mello.[5]

Parret (1988) trata da dependência que o ponto de vista da **pragmática linguística tem das noções semânticas** e considera que a modalidade seja, por excelência, parte de um estudo semântico. O autor comenta a ineficiência de análises puramente sintáticas: "é necessário recorrer a uma semântica, se não a uma pragmática **das modalidades**, se se quer recuperar a **estrutura distribucional**, mesmo que superficial, das modalidades" (PARRET, 1988, p. 83-84). Pessoa (2008) considera que afirmação de Parret (*op. cit.*) possua **fundamentação empírica**,[6] pois é fato que algumas formas modais estão distribuídas de modo diferentes de suas **variantes perifrásticas equivalentes**. Em seu trabalho, a autora aborda a utilidade dos mecanismos pragmáticos na **desambiguação** dos valores modais. Essa consideração é recorrente na literatura sobre o tema, tal como em Kärkkäinen (1989; 1992; 2003) e em Carretero (1992).

Pessoa (2008) assume que quando o falante usa uma forma modal canônica ele participa da enunciação, mostrando seu

[5] São eles, dentre outros:

Advérbios modalizadores na fala espontânea do português brasileiro: um estudo baseado em corpus (PQ CNPq Processo 311075/2009-6 - Professora Heliana Mello)

Modalidade na fala espontânea do Português Brasileiro: um estudo de corpus (FAPEMIG/UFMG Processo SHA-PPM 00324-08 - Professora Heliana Mello)

Valências modais em enunciados complexos (FAPEMIG/UFMG 1/2009 Aluna Priscila Côrtes)

O uso de advérbios modalizadores no português brasileiro (PIBIC CNPq 1/2009 Aluna Adriana Ramos)

Construções adverbiais modalizadoras na fala espontânea: um estudo prospectivo (PIBIC CNPq 1/2010 Aluna Raíssa Caetano)

Modalidade na fala espontânea: domínio de aplicação e a interface semântico-pragmática (PIBIC CNPq 1/2011 Aluna Raíssa Caetano)

[6] É importante notar que, muitas vezes, não se delimita o que se considera uma fundamentação empírica. A seguir, explicaremos a partir de qual viés tomamos nossos dados, ancorados no aparato metodológico da Linguística de Corpus.

engajamento no que diz e que, ao contrário, ao usar a **variante perifrástica**, este se ausenta da enunciação. O problema está justamente em definir o que chamamos de "variantes perifrásticas". Segundo as considerações de Pessoa (*op. cit.*), um índice canonizado e sua variante perifrástica são itens intercambiáveis semanticamente, apesar de se distribuírem de diferentes formas na cadeia sintática e provocarem efeitos diferentes na comunicação. Observamos aí a consideração de três níveis de análise: um semântico, no qual se daria a dita equivalência; um sintático, em que não haveria um pareamento, ou seja, há usos exclusivos em posições ou construções diferentes e um pragmático, no qual seriam percebidas alterações no sentido a depender da escolha do falante.

Castilho & Ilari (2008) assumem a existência de **"predicadores semelhantes"** ao comentar o mecanismo da paráfrase para analisar os advérbios modalizadores. Para os autores, em "realmente... [os filmes] eram muito ruins", *realmente* apresenta o conteúdo da sentença como um conhecimento válido. Afirma-se que o falante utiliza o advérbio para informar sua certeza, pois ele *sabe* que os filmes eram ruins, porém poderia ter utilizado outras formas de dizê-lo, conforme abaixo:

(3-1i) eu sei que os filmes eram muito ruins
(3-1ii) é certo que os filmes eram muito ruins
(3-1iii) é claro que os filme eram muito ruins
(3-1iv) na verdade, os filmes eram muito ruins

Em se tratando de fala, é necessário observar que esta não possui a mesma estruturação sintática da escrita. Cresti (2000), ao formular a Teoria da Língua em Ato (TLA), postula que a unidade de referência da fala seja o enunciado, definido como "a unidade mínima autônoma pragmaticamente". Para discutir tais questões de semelhança, devemos primeiramente revisar a análise do papel do nível semântico na comunicação falada. Isso se torna possível somente através de um *corpus* oral com um desenho adequado (cf. BIBER, 1993), como o C-ORAL-BRASIL (RASO & MELLO, 2012). Amparando a arquitetura de compilação de tal *corpus*, encontra-se a TLA, que considera a interface prosódia-pragmática como indispensável ao estudo da fala espontânea. Muitos estudos sobre aspectos pragmáticos foram desenvolvidos nesse cenário (dentre eles, VALE, H. P, 2010; ROCHA, B. M. A, 2011; RASO, T.; GOULART, L. L., 2009).

Todavia, as investigações sobre o nível semântico ainda se apresentam como uma incógnita.

Uma questão muito discutida no estudo da semântica linguística é a da *sinonímia:* até que ponto diferentes palavras e expressões seriam sinônimas e qual seria o papel do contexto linguístico e situacional nessa decisão. Biber *et al*. (1998) realizam um estudo baseado em *corpus* com o intuito de estabelecer uma distinção entre quase-sinônimos, partindo da Teoria Contextual da Sinonímia (baseada em WITTGENSTEIN, 1953 e FIRTH, 1957). Concordamos com os autores no que concerne à necessidade de que tais questões sejam discutidas a partir da análise de dados empíricos. Para tanto, realizamos um estudo baseado em um *corpus* de fala espontânea, como já mencionado, o qual será mais bem explicado na seção 2. A partir das novas metodologias encontradas para os estudos linguísticos, como aquelas oferecidas pela Linguística de *Corpus*, podemos discutir em pormenores a relação entre aquilo que regula o discurso (mutável) e aquilo que o constitui (fundamental).

Retornemos à questão da modalidade e suas diversas formas de expressão. García (2000) comenta a variedade e sofisticação da expressão da modalidade, e, como exemplo, apresenta uma lista de índices modais. O autor defende que a expressão da modalidade não esteja restrita a verbos modais, visto que a categoria pode se expressar de diversas formas "não-listáveis". Para complementar suas considerações, cita Halliday (1970), que postula que a modalidade não se situa em um só lugar na cláusula, o que é também discutido por Tucci (2007) a partir de um estudo de fala. Halliday (*op. cit*.) ressalta a recorrência de inúmeras "combinações" em uma mesma parte do discurso, isto é, possibilidade de diferentes formas de **distribuição estrutural**. Para Halliday

> ***nem as diferentes formas não-verbais do mesmo item lexical correspondem necessariamente umas com as outras: "obviamente" não é o mesmo que "é evidente que ..."," certamente " não é equivalente a "tenho certeza de que "***. Contudo, há grupos discerníveis, e uma clara distinção pode ser estabelecida entre pares que sejam considerados ***equivalentes,*** e, portanto, ***reforçam-se mutuamente***, quando ambos estão presentes, como em "Talvez ele possa ter construído", e aqueles que ***não são equivalentes*** e, portanto, em ***sentido cumulativo***, como em

> "Certamente, ele poderia ter construído" ("Eu insisto que é possível" ou "Admito que é possível") [grifos nossos].[7]

O autor comenta as noções de reforço, acumulação e equivalência de sentido que, segundo García (2000). Cresti (2002) e Tucci (2007), ao discutirem as questões de reforço e acumulação, defendem que quando há mais de um índice modal em um enunciado não é possível observar *composicionalidade*. Isso se dá porque eles têm seus escopos em diferentes unidades informacionais. A unidade informacional seria, assim, considerada como a "unidade de status neutro", o *dictum*, sobre o qual o índice modal opera, o *modus*. Dessa forma, poderíamos dizer que a unidade informacional é o escopo da modalidade. Neste trabalho, nos limitamos a discutir a noção de equivalência semântica a partir da observação da ocorrência dos pares modais, além de verificar o comportamento dos índices também nos níveis sintático e pragmático.

3 Metodologia

Como foi mencionado na seção anterior, em se tratando do posicionamento teórico para a investigação da expressão da modalidade na interface semântica-pragmática, consideramos que essa relação deva ser pesquisada em dados de fala espontânea. Para cumprir tal objetivo, pesquisamos no *corpus* C-ORAL-BRASIL[8] (RASO & MELLO, 2012). Esse *corpus* é uma ramificação do projeto C-ORAL-ROM (CRESTI & MONEGLIA, 2005), que reúne *corpora* de fala espontânea das principais línguas românicas da Europa, a saber, espanhol, francês, italiano e português europeu.

---

[7] Tradução nossa. "Nor do the different non-verbal forms of the same lexical item necessarily correspond with each other: "obviously" is not the same as "it is obvious that...", "surely" as "I am sure that". But there are discernible groupings, and a clear distinction can be drawn between pairs which are felt to be equivalent, and thus reinforce each other ("as concord") when both are present, as in "Perhaps he might have built it", and those which are not equivalent and are thus cumulative in meaning, as in "Certainly he might have built it" ("I insist that it is possible" or "I grant that it is possible")."

[8] Para mais informações, cf. <http://c-oral-brasil.org/>.

Ambos os *corpora*, foram arquitetados segundo a Teoria da Língua em Ato (CRESTI, 2000), que afirma que o enunciado é a menor unidade linguística possível de ser interpretada pragmaticamente. Todo falante de uma língua qualquer é capaz de perceber quebras prosódicas terminais e não-terminais no discurso oral. Essas quebras marcariam unidades informacionais, no nível pragmático, e unidades tonais, no nível acústico. Uma unidade percebida como terminal marcaria o fim de um enunciado. Assim, o *corpus* é segmentado prosodicamente e etiquetado informacionalmente, além de apresentar um alinhamento texto-som das transcrições e dos arquivos de áudio contendo cada gravação.

O *corpus* C-ORAL-BRASIL tem por objetivo representar a variação diafásica, sobretudo do dialeto mineiro do Português Brasileiro. O *corpus* se divide nos registros formal e informal.[9] A parte informal do corpus já está disponível, contendo 208.130 palavras, num total de 21:08:52 horas. Além de tal divisão, também se leva em consideração os contextos familiar/privado e público Contamos com a especificação de três tipos textuais, a saber, monólogos, diálogos e conversações, a depender dos números de participantes envolvidos na interação. Os transcrições são acompanhadas por cabeçalhos que contêm informações extra-textuais, que especificam as características dos falantes (idade, sexo, nível de instrução, profissão e naturalidade) e as particularidades da situação.

Os textos do corpus contam com dois tipos de etiquetagem, uma informacional e outra morfossintática. A etiquetagem informacional informa as diferentes funções desempenhadas pelas unidades informacionais (correspondentes pragmáticas da unidade tonal), individualizadas a partir de critérios funcionais, distribucionais e prosódicos. A unidade básica é a de COM – comentário, ou seja, todo enunciado deve conter pelo menos essa unidade, pois é ela a responsável por carregar a força ilocucionária. Atualmente, somente o minicorpus C-ORAL-BRASIL, que é uma parcela representativa desse tanto em aspectos formais quantos funcionais, possui a etiquetagem informacional. Isso se deve ao fato de que tal anotação é feita manualmente e carece de muitas revisões para que sua credibilidade seja assegurada.

[9] Vale salientar que pesquisamos somente na parte informal do *corpus*, visto que a parte formal se encontra em construção.

Já a etiquetagem morfossintática é realizada automaticamente pelo parser PALAVRAS (BICK, 2012). O anotador foi desenvolvido especialmente para anotação de PoS-tagging em língua portuguesa. Além disso, o parser teve de ser alimentado com regras que compreendem especificidades da fala espontânea, representadas no corpus através da transcrição. As transcrições são semi-ortográficas, por exemplo, a variação entre *cê*, *ocê* e *você* é preservada. A decisão foi tomada considerando-se a necessidade de preservar a produção real do falante, apesar de o acesso ao áudio ser possível, já que esta pode indicar processos de alteração no sentido e função de dado termo.

Foi feito um cotejamento dos itens considerados nesta pesquisa entre os dados encontrados no *corpus* C-ORAL-BRASIL e os do *corpus* de referência *Corpus* do Português (DAVIES & FERREIRA, 2006). O objetivo de se utilizar um *corpus* de referência na pesquisa é assegurar a confiabilidade da análise dos dados, visto que os *corpora* de referência "pretendem representar uma dada língua como um todo" (MELLO, 2012, p. 32). O *Corpus* do Português se constitui de 57000 textos, contendo 45 milhões palavras, compreendendo o período dos séculos XIV ao XX. Ele se divide nas variantes Português Brasileiro e Português Europeu, nos registros acadêmico, ficcional, jornalístico. A busca dos itens no C-ORAL-BRASIL foi mediada pelo uso do software TextSTAT, que organiza a frequência de palavras em texto e possibilita a visualização do contexto das mesmas.[10] Já no *Corpus* do Português foi utilizada a plataforma de busca que o *site* oferece, selecionamos a seção oral do português brasileiro, para assegurar a comparabilidade entre os corpora.

Há que salientar que a observação e análise dos dados estão condicionadas à estrutura que cada *corpus* apresenta. Por exemplo, as transcrições do C-ORAL-BRASIL respeitam a segmentação prosódica da fala e certos fenômenos fônicos que a língua sofre, já as transcrições da parte oral do *Corpus* do Português são baseadas em outros critérios, o que pode tornar difícil até mesmo a análise sintática do enunciado, que pode ser segmentado de várias formas somente com a transcrição, sem conferir a prosódia (cf. MELLO, 2012, p. 37). No *Corpus* do Português os arquivos de áudio da parte oral não estão disponíveis para consulta, assim sendo, as características orais da

[10] http://neon.niederlandistik.fu-berlin.de/en/textstat/

língua portuguesa só podem ser analisadas através das transcrições. Mello (2012, p. 33) já atenta para o fato de que "a produção de *corpora* eletrônicos orais no Brasil, entretanto, ainda é bastante restrita e necessita ser fomentada".

Realizada a busca nos *corpora*, os dados foram quantificados (cf. Tabela 1, seção 3) e passou-se a observar se a frequência dos itens de cada par era proporcional. Em uma segunda etapa, analisamos qualitativamente as ocorrências dos advérbios *logicamente, claramente, obviamente* e seus pares a partir do corpus C-ORAL-BRASIL e fizemos uma investigação diacrônica no Corpus do Português a respeito do advérbio *realmente*. A análise dos dados pesquisado nos dois *corpora* é apresentada na próxima seção.

4 Análise e resultados

A análise dos dados desenvolveu-se em dois momentos, o primeiro de cunho quantitativo e o seguinte de cunho qualitativo. No momento inicial, buscamos observar contrastivamente a ocorrência dos marcadores "realmente/na realidade, obviamente/é óbvio, claramente/é claro e logicamente/é lógico".[11] Na tabela a seguir é mostrada a quantidade de ocorrências dos itens investigados encontrada nos dois *corpora*:

[11] O *corpus* de referência *Corpus* do Português não foi utilizado em uma primeira análise, na qual ainda contávamos com o par *realmente/na realidade*. O cotejamento havia sido feito com o outro *minicorpus* composto por entrevistas coletadas na *internet*.

TABELA 1
Frequência dos pares modais nos dois *corpora* pesquisados[12]

Índices modais	**Ocorrências nos dois corpora**	
	C-ORAL-BRASIL	**Corpus de referência (CdP)**
claramente	1	95
(é) claro (que)	65	1306
logicamente	2	6
(é) lógico (que)	35	48
obviamente	4	119
(é) óbvio (que)	3	65
realmente[13]	32	1000
na realidade	8	100

Pode-se perceber que as frequências entre os dois *corpora* são proporcionais, o que assegura nossa análise. Os resultados apontaram para uma diferença quantitativa entre as variantes: *(é) claro (que)* e *(é) lógico (que)* ocorrem em maior proporção do que *claramente* e *logicamente*; *obviamente* e *realmente* ocorrem mais do que *(é) óbvio (que)* e *na realidade*. Tal análise seria um indício da não equivalência pragmática entre os índices, isto é, o estudo de frequências aponta para a preferência no uso de um índice. Propomos, então, uma segunda discussão, para averiguarmos também os âmbitos sintático e semântico, considerando o estudo da *prosódia semântica*. O termo refere-se "à associação recorrente entre itens lexicais e um campo semântico, indicando uma certa conotação (negativa, positiva ou neutra) ou instância avaliativa" (SARDINHA, 2004). Com relação à camada pragmática, observamos a distribuição por tipos textuais (conversação, diálogo, monólogo) no *corpus* C-ORAL-BRASIL. Em

[12] A contagem de dados aparece na forma simples e não em frequência relativa devido às amostras serem muito grandes frentes às ocorrências.

[13] Os valores do par *realmente/na realidade* estão descritos na tabela, mas, como já comentado, não iremos discuti-los contrastivamente. Pode-se observar que a frequência do advérbio *realmente* se apresenta sempre como a mais alta, como comentaremos à frente.

análise mais pontual, isolamos o estudo do advérbio *realmente* dos demais, devido à peculiaridade apresentada por esse índice.

Iniciemos com a descrição dos comportamentos semântico, sintático e pragmático assumidos pelos demais índices em análise. As variantes "claro/claramente" e "lógico/logicamente" apresentaram as mesmas distinções. Com relação a aspectos semânticos, todos os índices apresentam uma prosódia semântica neutra. Em âmbito sintático, *claramente* e *logicamente* ocorrem junto à proposição. É importante lembrar que a análise sintática se dá dentro da unidade informacional completa e não diz respeito às relações estabelecidas entre as mesmas. Confiram-se os exemplos abaixo:

> *LUC: mas / por exemplo / o Van Gogh / ele [/1] &c [/1] &e [/1] **claramente** / a [/1] as pinceladas / são muito importantes pra ele //$ *bfamdl09*[14]

> *JOR: com as amizades adquirida / que nós chamamos de "network" / &he / me apareceu uma outra / hhh oportunidade dentro de uma outra multinacional / aonde eu fui desenvolver / um trabalho de vendas / &he / junto / ao mercado / concorrente dessa empresa onde eu estava / e lá eu fiquei um período / desenvolvendo o mesmo tipo de trabalho / **logicamente** com um salário melhor / hhh e por amizade eu fui cair / em uma multinacional / que eu dei uma virada no produto //$ *bfammn06*

Já *claro* e *lógico* preenchem todo o conteúdo locutivo de um enunciado, como podemos observar nos exemplos seguintes. Esses índices, geralmente, são produzidos por um falante diferente daquele que afirmou o conteúdo a ser modalizado.

[14] A sigla *bfamdlX* identifica um texto do C-ORAL-BRASIL, e significa: *corpus* 'Brasil', contexto 'familiar', número de falantes 'diálogo' e X, o número do texto. *LUC indica as iniciais do nome do falante. As barras simples marcam as quebras prosódicas percebidas como não-terminais, enquanto que as barras duplas marcam quebras prosódicas percebidas como terminais, marcando o fim de um enunciado.

*DFL: <muito caladão> // e ele / brincalhão / porque era a única <filha> / né //
*LUC: <ham ham> // <**claro**> // *bfammn02*

*FLA: <oh Lud / deixa a primeira pra eles mesmo> porque / <se não nu vai dar muita pressão> //*EME: <**lógico**>// *bfamcv21*

Em termos pragmáticos, os dados mostram que *claro* e *lógico* são utilizados em interações dialógicas. Nesses contextos podem atuar como reguladores discursivos, atuando em uma camada pragmática, sem escopo semântico preciso. Tal aspecto deve ser melhor investigado, o que permite uma discussão mais apurada sobre a interface semântica/pragmática.

O par "óbvio/obviamente" apresentou comportamento similar, talvez por ambos possuírem uma prosódia semântica negativa, o que pôde ser mais bem observado nas ocorrências encontradas no *Corpus* do Português. Confira-se, abaixo:

(1) Não que não venhamos a cometer alguns **erros**, mas se podermos evitar alguns erros com aquilo que já sabemos, com o passado, *é* ***óbvio*** *que* convém fazê-lo.

(2) *É* ***óbvio*** *que* existe **dificuldades**, mas não tanto assim.

(3) Não estou, como *é* ***óbvio***, **nada de acordo** com as declarações de Manuela Morgado. Além disso, é, no mínimo, **desagradável** que depois de se demitir ela teça essas considerações contra o Governo.
CdP (DAVIES & FERREIRA, 2006)

A correspondência em âmbito pragmático seria observada caso consideremos que os índices concedem maior grau de certeza à asserção. Os índices são equivalentes também em âmbito sintático, por ocorrem no interior de um único enunciado, como podemos perceber através dos exemplos retirados do C-ORAL-BRASIL:

*TUT: <sem ela saber> / e foi pegando ela //
*CLA: <obviamente> // *bfamcv30*

*BRU: tá falando do meu pé / né //
*CEL: óbvio // *bfamcv04*

Como podemos observar a partir da análise dos dados, contribuições importantes para a melhor compreensão da modalidade em línguas naturais podem ser alcançadas através de estudos de *corpora*, como propomos, uma vez que assim observam-se tendências estatisticamente significantes de padrões e índices modalizadores, como feito por Mello & Caetano (em preparação) no mapeamento de advérbios modais. Além disso, podemos analisar a língua como instrumento de interação social e verificar seus componentes estruturais. Na fala espontânea, é necessário estarmos atentos às faces ilocutiva e locutiva que compõem os atos de fala, para depois expandir os horizontes das análises sintática e semântica.

Passemos agora à análise do índice *realmente* e seus possíveis correlatos. O primeiro impasse com relação ao estudo desse advérbio foi a dificuldade em encontrar um correlato, visto que não temos a correspondência que há entre os outros pares, como em *claramente* e *(é) claro (que)*, por exemplo. Inicialmente, consideramos a locução *na realidade*, mas a análise contrastiva não apresentou os mesmos aspectos observados na oposição entre os demais pares. Dessa forma, retornamos à construção correspondente nos outros casos, que seria *(é) real (que)*, para investigá-la diacronicamente, a partir do *Corpus* do Português. A construção poderia não ter sido afetada por um processo de gramaticalização em direção ao sentido modal tal como as demais.

Na análise de dados, encontramos ocorrências em que a construção é utilizada em co-texto modal, o que não se mostrou significativo quantitativamente. Abaixo, uma ocorrência retirada do *Corpus* do Português:

(4) P. – Que dados concretos é que o embaixador Butler tem para afirmar que o Iraque **pode** construir armas biológicas numa semana?

R. – Perguntamos a ele isso e o que ele disse é que esse é um 'prazo **possível'**. Ele não tem nenhumas provas concretas.

R. – Não é uma afirmação alarmista?

R. – **É verdade que** foi uma afirmação não baseada em dados concretos. Mas tecnicamente é **possível**.

> P. – Até que ponto ***é real*** *a hipótese* de uma ação militar?
>
> R. – **Temos que** esgotar todos os meios diplomáticos, mas **é evidente que**, **se** houver um ato de agressão iraquiana, isso altera a situação.
>
> CdP (DAVIES & FERREIRA, 2006)

Seria possível que a atribuição de um sentido modal a tal forma não tenha vingado no uso da língua. Voltamo-nos, então, para a análise acurada do advérbio *realmente*. Os advérbios com terminação em *–mente* já são fruto de gramaticalização. A formação destes se deu a partir da adição da palavra *mente*, anteriormente utilizada em seu sentido lexical. Mello & Caetano (em preparação) constatam, em análise quantitativa, que advérbios modais desse tipo apresentam uma alta produtividade morfológica. Baayen (2009) define esse fenômeno com base em estudos de *corpus*. Para o autor, morfemas produtivos seriam aqueles que apresentam uma alta produtividade de tipos com baixa frequência de *tokens*. No mapeamento feito por Mello & Caetano (em preparação), do total de 763 ocorrências de advérbios e construções adverbiais presentes no *corpus* C-ORAL-BRASIL, esses advérbios modais representam apenas 46 ocorrências. Contudo, de 28 tipos encontrados no *corpus*, 18 deles são advérbios terminados em *–mente*, o que caracteriza a produtividade de advérbios com tal formação.[15] É interessante notar que o advérbio *realmente* é o único a apresentar uma alta frequência, com 31 ocorrências.

Dessa forma, nos perguntamos se existiria algum fator que haveria condicionado o diferente emprego dos advérbios gramaticalizados. O advérbio *realmente* parece estar em um novo processo de gramaticalização. Com relação à construção *(é) real (que)*, intuímos que *real* possua um valor semântico mais forte do que *claro* e *lógico* e por tal motivo tenha resistido a um processo de gramaticalização. O termo *claro*, por exemplo, também é utilizado como adjetivo em oposição a *escuro*, além de estar correlacionado a "esclarecer". Da mesma forma, *lógico* funciona tanto como substantivo

[15] Os advérbios modais terminados em *–mente* são: *realmente, exatamente, justamente, provavelmente, sinceramente, necessariamente, obviamente, evidentemente, logicamente, possivelmente, certamente, claramente, fatalmente, aparentemente, terminantemente, definitivamente, eventualmente* e *potencialmente*.

e adjetivo (matemática). Por outro lado, *real* exerce somente função de adjetivo, *real VS fictício*, a qual se relaciona com a noção de factualidade, que, por sua vez, dialoga com a noção de modalidade, questão discutida na literatura sobre o fenômeno (cf. SWEETSER, 1990).

5 Considerações finais

A partir da análise de dados, podemos afirmar que aspectos pragmáticos estão relacionados com o uso dos modais, mas há que se encontrar o limite tênue para a análise semântica dos marcadores. Os modais são empregados com diferentes graus de certeza de acordo com o tipo de relação estabelecido entre os participantes, constatação que carece de mais observação do uso para se solidificar. Além disso, os modais passam a assumir funções inteiramente pragmáticas quando atuam como reguladores do discurso, a depender do tipo de interação. O estudo mais aprimorado da interface semântico-pragmática é desenvolvido por Mello & Caetano (em preparação) em uma investigação de cunho qualitativo. Em monólogos, apresentam-se com um escopo semântico esparso, o que leva à reanálise de seu sentido semântico.

A investigação em *corpus* responde à nossa questão inicial, apontando para a não equivalência total dos índices em diferentes níveis da língua. Os pares não parecem ser correlatos, de modo que não sejam concorrentes. Contudo, podem ser coocorrentes na língua, como revelado pelo estudo de "óbvio/obviamente" (prosódia semântica). No entanto, essa variação não é livre, consideramos que haja um condicionamento sintático, semântico e pragmático. A partir de tal estudo embrionário, podemos reafirmar a necessidade de revermos nossas conceptualizações a respeito dos diferentes níveis da língua a partir de estudos da fala espontânea, com base em *corpora* adequadamente estruturados.

Agradecimento

Agradecemos à professora Heliana Mello, não só pelo auxilio na elaboração deste artigo e no desenvolvimento de outras pesquisas, como também por ser exemplo de postura.

Referências

BAAYEN, P. H. Corpus linguistics in morphology: Morphological productivity. In: LUDELING, A.; KYTO, M. (Ed.). *Corpus linguistics*: An international handbook. Vol. 2. Berlin: Mouton de Gruyter, 2009.

BIBER, D. Representativeness in corpus design. *Literary and Linguistic Computing* 8. p. 243-257, 1993.

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

BICK, E. C-ORAL-BRASIL grammatical tagging. In: RASO, T.; MELLO, H (Org.). *C-ORAL-BRASIL I*: Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012.

BYBEE, J. L. *Morphology:* a study of the relation between meaning and form. Amsterdam: John Benjamins. 1985.

CARRETERO, M. The role of epistemic modality in English politeness strategies. *Miscelánea*, v. 13, p. 17-35, 1992.

CRESTI, E.; MONEGLIA, M. *C-ORAL-ROM*. Integrated reference corpora for spoken Romance languages. Amsterdam/Philadelphia: John Benjamins, 2005.

CRESTI, E. Illocuzione e modalità. In: BECCARIA, P.; MARELLO, C. (a cura di). *La parola al testo.* Scritti per Bice Mortara-Garavelli, Ed. dell'Orso, Torino. 2001. p. 133-145.

CRESTI, E. *Enunciato e frase*. Firenze: Academia della Crusca, 2000.

DAVIES, M.; FERREIRA, M. *Corpus do Português.* Disponível em: <http://www.corpusdoportugues.org/>. Acesso em: 29/10/2011.

GARCÍA, F. G. *Modulating grammar through modality:* a discourse approach. Universidad de Almería. ELIA I, 2000.

GASPARINI-BASTOS, S. D. *Uma descrição do comportamento dos advérbios modalizadores epistêmicos no português falado*. Dissertação (Mestrado). Campinas: UNICAMP, 1997.

HALLIDAY, M A. K. Functional Diversity in Language as Seen from a Consideration of Modality and Mood in English. *Foundations of Language* 6: p. 322-361, 1970.

ILARI, R.; NEVES, M. H. M.; CASTILHO, A. T. (Org.). *Gramática do Português Culto Falado no Brasil*. 1. ed. Campinas: Editora da Unicamp, 2008. v. 1. 1167 p.

KÄRKKÄINEN, E. On the functions of epistemic modality in English discourse. In: JÄNTTI, A. (Ed.) *Probleme der Modalität in der Sprachforschung*. Studia Philologica Jyväskyläensia 23, Universität Jyväskylä, Jyväskylä 1989. p. 149-158, 1989.

KÄRKKÄINEN, E. Modality as a strategy in interaction: Epistemic modality in the language of native and non-native speakers of English. In: BOUTON, L.; KACHRU, Y. (Ed.). *Pragmatics and Language Learning.* Monograph Series volume 3, University of Illinois at Urbana-Champaign. p. 197-216, 1992.

KÄRKKÄINEN, E. *Epistemic Stance in English Conversation:* a description of its interactional functions, with a focus on *I think*. Amsterdam/Philadelphia: John Benjamins, 2003.

LYONS, J. *Semantics*. Vol. 2. Cambridge: Cambridge University Press, 1977.

MELLO, H. Os *corpora* orais e o C-ORAL-BRASIL. In: RASO, T.; MELLO, H (Org.). *C-ORAL-BRASIL I*: Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. p. 31-54.

MELLO, H.; CAETANO, R. V. O. *Mapeamento de construções adverbiais modais na fala espontânea:* um estudo baseado em corpus. Em preparação.

MELLO, H.; CAETANO, R. V. O. *Modalidade na fala espontânea:* um estudo sobre a interface semântico-pragmática. Em preparação.

PESSOA, N. P. A categoria modalidade e a (in)determinação de fronteiras. In: *I Simpósio Mundial de Estudos de Língua Portuguesa*, 2008, São Paulo. I Simpósio Mundial de Estudos de Língua Portuguesa, 2008.

PARRET, H. *Enunciação e Pragmática*. Trad. Eni P. Orlandi; Marco A. Escobar; Maria A. Babo; Paulo Otoni; Raquel S. Fiad e Rodolfo Ilari. Campinas: Editora da UNICAMP, 1988.

RASO, T.; MELLO, H (Org.). *C-ORAL-BRASIL I*: Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012.

RASO, T.; GOULART, L. L. As unidades informacionais de alocutivo em italiano e português do Brasil. *Fragmentos*, v. 9, p. 84-96, 2009.

ROCHA, B. M. A *Unidade Informacional de Introdutor Locutivo no Português Brasileiro:* uma análise baseada em corpus. Dissertação (Mestrado) Faculdade de Letras da Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, 2011. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/DAJR-8ELJXZ>.

SARDINHA, Tony Beber. *Lingüística de corpus*. Barueri: Manole, 2004.

SWEETSER, E. *From etymology to pragmatics*: metaphorical and cultural aspects of semantic structure. Cambridge: Cambridge University Press, 1990.

TUCCI, I. *The informational structure and the scope of lexical modality in spoken Italian*. Lablita, University of Florence, 2007.

VALE, H. P. *A unidade informacional de parentético no português do Brasil:* uma análise baseada em corpus. Dissertação (Mestrado) Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, UFMG, 2010.

Corpos e cores: colorindo a descrição da língua portuguesa

Cláudia Freitas[1]
Diana Santos[2]
Rosário Silva[3]

RESUMO: Na senda de nossos estudos anteriores, apresentamos aqui um quadro das diferenças de uso entre as cores nas variantes brasileira e portuguesa do português, com ênfase nos verbos. Após uma panorâmica dos recursos utilizados e da filosofia e prática de anotação subjacente, apresentamos as pesquisas sobre palavras de cor e seus argumentos, de uma forma didática e convidando os leitores a repetir as buscas ou a efetuarem outras no nosso material. No artigo apresentamos também o corte-e-costura, a ferramenta de apoio à anotação do projeto AC/DC, e o tipo de regras utilizadas. Focando a nossa atenção na classe gramatical dos verbos, concluímos que estes tendem a perder o seu significado originalmente colorido e a especializarem-se em sentidos muitas vezes metafóricos ou restritos a conotações ou mesmo expressões idiomáticas. É além disso nesse sentido, principalmente, que observamos diferentes usos nas duas variantes.

PALAVRAS-CHAVE: Linguística com corpos, cores, português, variantes do português, verbos

ABSTRACT: Following up previous studies on the subject of colours in Portuguese and possible differences between the two variants, we consider in detail the verbs conveying colour in this paper. After presenting the resources and the underlying annotation philosophy and practice, we submit to the reader a set of searches on colour

[1] Doutora, professora do Departamento de Letras da PUC-Rio e colaboradora da Linguateca. Email: claudiafreitas@gmail.com

[2] Doutora, professora associada do departamento de línguas, literaturas, e culturas europeias da Faculdade de Letras da Universidade de Oslo, líder da Linguateca. Email: d.s.m.santos@ilos.uio.no

[3] Mestre, colaboradora da Linguateca. Email: mrosariomsilva@sapo.pt

verbs and their arguments, showing how they can be replicated or modified for further study. We also briefly present corte-e-costura, the annotation tool in the AC/DC project and the corresponding rules. Our focus here being the verbs, we concluded that they tend to lose their originally visual import and specialize in metaphorical meanings, connotatively laden expressions, or even idioms, and that this trend is the main reason for differences among the two varieties investigated.

KEYWORDS: Corpus linguistics, colours, Portuguese, Portuguese varieties, verbs

1 Introdução

A questão das palavras de cor, além do seu interesse intrínseco, linguístico e filosófico (repare-se que além de ilustrarem a atribuição de propriedades (e não apenas ações ou objetos), são prototipicamente aplicadas a referentes concretos), tem uma particularidade que a torna especialmente interessante para estudos com corpos: são suficientemente comuns para serem estudadas quantitativamente. E além de comuns, ou talvez por isso, são encontradas em muitíssimas expressões mais ou menos convencionais, em que perdem o seu significado básico e exprimem outros valores bem afastados do campo visual, tal como atitude, conotações, assim como refletem (ou podem refletir) usos e artefatos antigos (*passadeira vermelha/tapete vermelho*, *papel pardo*, *páginas amarelas*).

Com efeito, do ponto de vista de frequência, alguns adjetivos de cor são dos adjetivos mais frequentes, mas interessa saber até que ponto o são porque a cor é fundamental na língua portuguesa, ou o são porque tomaram muitos mais objetivos e sentidos além do prototípico (como é o caso de palavras como *gente* ou *coisa*, que se tornaram genéricos, ou *pé* ou *mão* que estão no caminho da gramaticalização em locuções como *ao pé de* (variante de Portugal) ou *à mão*).

Por outro lado, é possível que o número distinto de expressões convencionais com uma dada palavra de cor ateste a idade dessa cor na língua, e/ou a produtividade do uso dessa forma, assim como é possível que ateste diferenças ou aumento de sentidos.

Gill Philip (2011) ilustra de forma muito clara todo o espetro de questões semânticas na língua, desde colocação, expressões

idiomáticas e metaforização até ao uso criativo da língua, usando a cor nos corpos.

Como será referido na nossa resenha sobre trabalhos anteriores, na secção 1.1, a cor tem também já uma longa história de tratamento linguístico e estudo, mesmo baseado em corpos. Neste trabalho, damos continuidade à exploração das cores em corpos da língua portuguesa, ao mesmo tempo que, aproveitando as possibilidades de pesquisa em corpos das variantes do Brasil e de Portugal, contrastamos também o uso nas duas variantes.

De maneira geral, este trabalho pode ser entendido como uma continuação de Santos et al. (2011), que, ao apresentar dados que sugerem um maior uso das cores na variante portuguesa, levanta questões sobre o uso das cores no Brasil e em Portugal.

Partindo de uma investigação sobre a distribuição das categorias gramaticais nas duas variantes, comparamos, em um primeiro momento, a distribuição das cores por categoria gramatical. Além disso, tirando proveito de corpos ricamente anotados com informação linguística, exploramos a distribuição das cores por categoria sintática (e por variante). O objetivo do presente trabalho é duplo: por um lado, buscamos perceber como se relacionam cor e categoria gramatical nas variantes PT e PB; por outro, tentamos prover uma descrição detalhada do comportamento dos verbos de cor.

1.1 A cor nos estudos linguísticos

O estudo das cores ocupa um papel importante no debate sobre universalismo e relativismo, interessando a diferentes áreas do conhecimento. Como possui aspectos biológicos e linguísticos, é natural que o campo das cores seja de especial relevância nos estudos da/sobre a linguagem (Deutscher, 2010). No entanto, como notam Santos et al. (2011), boa parte dos estudos sobre a cor, defendendo a universalidade da conceptualização da cor (Berlin & Kay, 1969; Rosch, 1975) ou, pelo contrário, refutando-a (Wierzbicka, 1990), tomam como base experiências com informantes. Como também aponta Lucy (1997), as pesquisas sobre as cores têm se concentrado na comparação entre as línguas, em um refinamento da tipologia e no reforço de argumentos de base biológica, enquanto relativamente pouco tem sido feito para melhorar a qualidade da descrição linguística. Saunders (2005), numa perspetiva filosófica, denuncia a base biológica

implícita de Berlin & Kay (1969). Nesse contexto, trabalhos com base em corpos podem oferecer um bom complemento no que se refere ao comportamento das cores nas línguas. Com relação ao português, além da exploração que vimos desenvolvendo no âmbito do AC/DC (Santos et al., 2011), do CorTrad (Teixeira et al., 2012) e do COMPARA (Frankenberg-Garcia & Santos, 2002) relativa à cor (Inácio et al., 2008, Silva et al., 2008, Santos et al., 2008, Santos et al., 2011), referimos também Biderman et al. (2007), que parte igualmente de corpos para explorar e descrever as diferenças na expressão da cor no Brasil e em Portugal. Ainda para o português, remetemos o leitor para os trabalhos de Bacelar do Nascimento et al. (1996), Correia (2006) e Farias & Marcuschi (2006), que se debruçam sobre diferentes questões relacionadas com a cor, e Jorge et al. (2003) e Zavaglia (2006, 2007), que tratam o fenômeno de uma perspectiva contrastiva e/ou lexicográfica – veja-se também Philip (2003) para o contraste entre o inglês e o italiano. Finalmente, numa ótica diacrônica, Sletsjøe (1962) e Swearingen (1990) dissertam sobre, respetivamente, o conceito de verde aplicado a olhos e o da diferenciação dos termos vermelho e roxo em português.

2 Contexto

Nossa exploração de corpos toma por base o projeto AC/DC (Acesso a Corpos / Disponibilização de Corpos), que disponibiliza corpos do português, nas variantes do Brasil e de Portugal, na Internet, e contém atualmente cerca de 374 milhões de palavras (Costa et al, 2009, Santos, no prelo). Todo o material foi anotado automaticamente pelo analisador sintático PALAVRAS (Bick, 2000), tendo algumas partes passado por revisão humana. Além da anotação do PALAVRAS, os corpos do AC/DC também vêm recebendo anotação relativa à informação semântica no campo das cores (Mota & Santos, 2009) como mencionado na secção anterior. A seção seguinte apresenta brevemente as opções linguísticas subjacentes à anotação das cores, bem como problematiza os casos mais difíceis. Para a documentação detalhada do trabalho de anotação linguística das cores, consulte-se o Arco-Íris (Silva e Santos, 2012).

2.1 A anotação semântica

Como mencionado, todos os corpos do AC/DC possuem informação semântica relativa à cor, presente no atributo "sema". Atualmente, são 7 as classes de cor (veja seção 2.2 sobre o processo de anotação):

QUADRO 1
Classes de cores no AC/DC

SEMA	Explicação	Exemplo
[sema="cor"]	Cor pura, representação de atributos visuais	sapatos **vermelhos**; nuvens **cinzentas**
[sema="cor:humana"]	cores que correspondem a atributos naturais humanos	cabelos **ruivos,** ou **louros**; **corar** de vergonha
[sema="cor:vinho"]	Cores associadas aos tipos de vinho	vinho **branco, tinto** do Douro; espumante **rosé**
[sema="cor:raça"]	Cores usadas para representar raça	os **pele-vermelha**; **brancos**, **negros** e **pardos**
[sema="cor:política"]	Cores associadas a partidos ou ideias políticas por metonímia	deputados **verdes**; socialismo **vermelho**
[sema="cor:equipa"]	Cores associadas a equipes (geralmente de futebol) por metonímia	torcida **alvinegra**; defesa **encarnada**
[sema="cor:original"]	Usos não convencionais; não representativos de atributos visuais (e que não se enquadram nos anteriores)	período **negro**; concordata **branca**; vida **cor de rosa**

A classe original é atribuída quando estamos diante de usos não relacionados a propriedades visuais – quando o vínculo com a propriedade colorida se perdeu, total ou parcialmente. Naturalmente se pode argumentar que nos usos de [sema="cor:política"] e [sema="cor:equipa"] o vínculo com a cor também pode estar distante, mas o uso da cor nesses domínios é sistemático o suficiente para nos permitir agrupá-los sob um rótulo. No caso de [sema="cor:original"]

(que admitimos talvez não ser o melhor nome, mas que provém do seguinte raciocínio "teve origem em cor"), agrupamos tanto (i) usos metafóricos de uma cor, separados ou englobados em expressões fixas (ficar *no* ***vermelho***; *dar* ***branco***); (ii) termos que originalmente contêm uma palavra de cor, mas que são considerados unidades, e que não se referem sobretudo a cor (*buraco* ***negro****; elefante* ***branco***); (iii) termos que, sendo ainda possivelmente descritivos, são usados metaforicamente (*sinal* ***verde;*** *levou um cartão* ***vermelho*** (? *encontrei um cartão* ***vermelho*** *na mesa com um poema* (que é considerado [sema="cor"]). Além disso, (iv) usamos [sema="cor:original"] quando estamos diante de usos não previstos, que podemos apelidar de criativos, das palavras de cor, como ilustram os trechos (1-4) a seguir:

(1) Se definem como socialistas (ou «socialismo **moreno**», como diz Brizola) , mas tem forte apelo populista .

(2) Ainda que tingido do ' socialismo **moreno**» do autor de "O Povo Brasileiro" , o livro ...

(3) Impressionado com o bronzeado dos colegas, o baiano Jaques Wagner (PT-BA) , desbotado pelas viagens à Alemanha, Coréia do Sul e Inglaterra, disse: ' O socialismo **moreno** do Brizola não deu certo, mas pelo visto o neoliberalismo **mulatinho** do Fernando Henrique vai bem».

(4) Sem a ajuda soviética e vítima do bloqueio americano, o sonho do socialismo **moreno**, que tanto encantou a esquerda brasileira, parecia estar fadado a um final infeliz.

Se *moreno* nos exemplos (1-4) é cor tanto quanto o seriam "socialismo vermelho" ou "socialismo verde", por exemplo, então poderíamos atribuir-lhe a etiqueta [sema="cor:política"]. No entanto, não temos a cor *morena* sistematicamente associada a nenhum campo da política em português, e portanto anotamos como [sema="cor:original"], o mesmo se aplicando a "mulatinho" na frase (3). (E notamos que a etiqueta [sema="cor:humana"] jamais seria atribuída nesses casos.)

Além das classes semânticas, as cores (quando sema="cor") também são classificadas quanto ao grupo:

Grupos de cor: BRANCO, PRETO, AZUL, AMARELO, VERMELHO, LARANJA, VERDE, ROXO, CASTANHO, CREME, CINZENTO, ROSA, PRATEADO, DOURADO e também OUTRAS.*, MÚLTIPLA.*, AUSÊNCIA, NÃOESPECIFICADA e DESCONHECIDA.

O grupo Outras engloba expressões cuja cor pareceu impossível identificar ou definir, tais como *cor de apoplexia*; *cor de crime e traição*; *cor de ferro velho*, *cor de morte* etc. Além disso, usamos também o grupo Outras quando não é possível decidir sobre (e portanto anotar) a inclusão de uma cor em um determinado grupo. *Pardo*, por exemplo, é castanho/marrom ou cinzento? E *cor de telha*?

É importante esclarecer que a relação entre os grupos e as classes de cor não é de complementação, e não há dupla categorização. A classificação de grupo de cor só existe se a ocorrência de cor for do tipo [sema= "cor"], isto é, se estivermos diante de usos descritivos, coloridos. Dessa forma, em *sorriso* ***amarelo***, o amarelo não pertence ao grupo Amarelo, visto ser [sema="cor:original"], e o mesmo para *vinho* ***verde***, classificado como [sema="cor:vinho"]. Estamos conscientes, no entanto, de que algumas das classes também são descritivas, como [sema="cor:humana"] (em *olhos* ***verdes*****,** não há como negar a propriedade colorida de "verde"), e uma possibilidade para o futuro é adicionar grupos de cores específicas para a classe humana, por exemplo.

As motivações para a existência do grupo atual de cores (com 19 elementos) são, sobretudo, empíricas. Tomamos como critério o uso variado e frequente de um termo para uma cor/grupo de cor, pois isto significa que, de alguma maneira, esse grupo/cor é relevante na língua. No grupo *Creme*, por exemplo, são 18 as instâncias de cores, como bege; bege-areia; creme; marfim; perolado; pérola etc., o que o torna merecedor de um grupo próprio. Além disso, a busca por um "grupo" é mais prática que uma busca por lema (vide os 18 lemas de *Creme*). Vale lembrar, no entanto, que o grupo não foi fixo *a priori*, e que a exploração das cores nos corpos pode levar à criação de novas categorias/grupos de cores.

2.2 A ferramenta de anotação (corte-e-costura)

O corte-e-costura (Mota & Santos, 2011, Santos & Mota, 2012) é uma ferramenta desenhada para ajudar a anotação dos corpos, a partir do desenvolvimento de regras sucessivamente mais específicas, para obter cobertura total, pondo o anotador no ciclo: ou seja, desenvolvendo um conjunto inicial de regras, aplicando-as, revendo o resultado, e continuando num processo iterativo até considerar toda a anotação correta, e abrangendo todos os casos.

É baseado num léxico, quer de palavras simples quer de expressões com várias palavras, e na aplicação de cinco tipos de regras, que em seguida exemplificamos para a cor: 1) regras positivas adicionando casos de cor que não estão no léxico, 2) regras negativas removendo casos que não deviam ser considerados cor; 3) regras de especialização que transformam casos de cor noutros mais específicos; 4) regras que retiram a marcação de casos específicos; e 5) regras recursivas (cuja aplicação só pára se não puderem continuar a ser aplicadas). Além disso, para cada tipo de regras existe a distinção entre regras gerais e regras específicas associadas a cada corpo, que fica ao critério do anotador, conhecendo o conjunto dos corpos do AC/DC, decidir.

Abaixo temos um exemplo de uma regra de cada tipo: 1) a palavra *celeste* é considerada cor se antecedida pelas palavras *bicicleta*, *divã* ou *vestimenta*; 2) a palavra *louro* não é palavra de cor quando na expressão *louros da vitória*; 3) a palavra *branco* refere-se a vinho se estiver seguida por *tinto*, e passa pois de simples cor a cor:vinho; 4) a palavra tricolor que tinha sido atribuída à classe equipa volta a ser apenas "cor" (no caso específico de um dado corpo); e 5) se as palavras *aço*, *banana*, *borgonha* e *café* no singular estiverem seguidas da conjunção *e* e de uma palavra de cor, são consideradas cor.

```
a:[lema="bicicleta|divã|vestimenta"] b:[lema="celeste"] >> b:[sema="cor"]

a:[word="louros"] b:[word="da|do"] c:[word="triunfo|vitoria"] >>
a:[sema="0"]

a:[lema="branco"] b:[word=","] c:[lema="tinto"] >> a:[sema="cor:vinho"]

[sema="cor:equipa" & lema="tricolor"] >> [sema="cor"]

a:[lema="aço|banana|borgonha|café" & pessnum="S"] b:[lema=",|e|ou"]
c:[sema="cor"] >> a:[sema="cor"]
```

3 Exploração das cores nos corpos

No presente trabalho, tomamos por base principalmente os corpos CONDIVport (Silva, 2008) e CHAVE (Rocha & Santos, 2007).

O CONDIVport, criado com o objetivo de permitir o estudo da convergência e divergência do português entre as variantes do Brasil e de Portugal, contém cerca de 5 milhões de palavras distribuídas em jornais desportivos e revistas de saúde e de moda – áreas que tendem a empregar cores de maneira recorrente. Além disso, é importante mencionar que toda a anotação das cores no CONDIV passou por revisão humana.

O CHAVE contém cerca de 99 milhões de palavras distribuídas em textos jornalísticos da Folha de São Paulo (Brasil) e do jornal Público (Portugal). Por ser um corpo maior, e de conteúdo mais geral, oferece dados complementares aos obtidos no CONDIVport, mas ainda não foi objeto de revisão humana.

3.1 Categorias gramaticais

A tabela 1 apresenta a distribuição das cores – sem distinção de tipo – por categoria gramatical nos corpos CHAVE e CONDIV. Como é possível observar, no CHAVE, não há qualquer diferença na distribuição das cores por categoria gramatical entre as variantes BR e PT, com uma ocorrência maior de adjetivos coloridos se comparados com os verbos. No CONDIV, os dados relacionados aos verbos coloridos também são idênticos entre as variantes, e por sua vez são menos frequentes que os verbos no CHAVE. Com relação aos adjetivos de cor, a situação no CONDIV é um pouco diferente, havendo mais adjetivos coloridos na variante PT do que na variante BR. Tal fato, no entanto, conforme explicado em Santos et al., (2011), pouco tem a ver com a distribuição de adjetivos, ou com o uso das cores, mas antes se deve à composição do corpo CONDIV: a porção BR do corpo contém muitos moldes de cor – e menos texto, em um sentido estrito – o que explica a diferença nos números.

TABELA 1
Distribuição das cores por categoria gramatical e variante

	CHAVE		CONDIV	
	ADJ cor	Verbos cor	ADJ cor	Verbos cor
BR	0,6%	0,2%	2,2%	0,06%
PT	0,6%	0,2%	3,4%	0,06%

3.2 Exploração dos verbos coloridos

Com relação aos verbos coloridos, nosso principal interesse nesse trabalho, constatamos que a imensa maioria está na forma de particípio, o que não surpreende, dado que o particípio frequentemente assume o papel de modificador. No entanto, como as formas participiais ainda não foram totalmente revistas no CHAVE (Douro (rio e região), e dourado (peixe), por exemplo, ainda estão como particípios passados do verbo *dourar*) optamos por não considerá-las neste trabalho.

A tabela 2 contém todos os verbos coloridos, excluindo o particípio, contidos no AC/DC,[4] em ambas as variantes. Os lemas marcados com ** ocorrem no CHAVE e no CONDIV, e portanto serão considerados no presente estudo. Também na tabela 2 indicamos a distribuição dos lemas por variante, independente de corpo, e notamos que, em termos gerais, há um uso mais frequente de verbos de cor na variante de Portugal. Embora não esteja na tabela, mas seja de fácil aferição na página do AC/DC, percebemos que os verbos presentes apenas na variante PT vêm, em grande parte, do corpo Vercial, que contém obras literárias de autores portugueses, cujas datas de publicação variam desde 1500 a 1933.[5]

[4] À data do presente trabalho. Como a anotação e sua revisão ainda estão em progresso, poderemos ainda encontrar mais casos no futuro, assim como modificar os dados quantitativos.

[5] O corpo Vercial contém obras digitalizadas pelo projeto Vercial, http://alfarrabio.di.uminho.pt/vercial/, e de fato podemos considerar que, exceto nos séculos XIX e XX, não faz sentido distinguir os autores "portugueses" dos "brasileiros". Um estudo mais fino deverá ser feito, reduzindo os escritores portugueses ao período moderno, e não entrando em conta com os outros, ou considerando alguns autores antigos como pertencendo às duas ou mesmo à variante brasileira, como poderia ser o caso do Padre António Vieira ou do Judeu.

TABELA 2
Distribuição dos verbos coloridos nos corpos CHAVE e CONDIV, sem particípio.[6]

Grupo de cor	Lema e distribuição dos verbos por corpo e variante					TOTAL por variante (de todos os corpos)	
		CHAVE		CONDIV			
		BR	PT	BR	PT	PT	BR
Grupo Vermelho	avermelhar**; vermelhejar; vermelhar, purpurar, revermelhar, purpurejar, enrubescer	4	4	2	10	50	14
Grupo Laranja	alaranjar**	1	6	—	—	17	2
Grupo Amarelo	amarelar**, amarelecer**, amarelejar	7 2	1 14	1 —	1 1	14 67	16 3
Grupo Roxo	arroxear**, roxear	1	1	—	—	2	—
Grupo Preto	enegrecer**, negrejar	2	13	—	4	**120**	6
Grupo Verde	esverdear**; reverdecer**; verdejar**, esverdear; esverdinhar; verdecer	— 1	8 1	— 1	1 —	25 69	3 6
Grupo Cinzento	acinzentar**	4	6	—	—	22	5
Grupo Rosa	rosar**, rosear	—	2	—	—	—	2
Grupo Dourado	dourar**/doirar**, sobredourar, sobre-dourar; sobredoirar	51	50	4	3	**857**	64

Continua

[6] Para realizar a busca por lema no AC/DC, após a escolha do corpo desejado para a pesquisa, a expressão de busca é (o exemplo refere-se à busca pelos verbos em forma finita pertencentes ao grupo Vermelho): [sema="cor.*" & grupo="Vermelho" & temcagr!="PCP" & pos="V.*"]. Para a distribuição por corpo, uma vez selecionado o corpo TODOS, a expressão de busca é (o exemplo refere-se à busca da distribuição, pelos corpos, do lema "branquear", na variante PT):

[sema="cor.*" & lema="branquear" & temcagr!="PCP" & pos="V.*" & variante="PT"], e nos resultados deve-se pedir "Distribuição de corpo".

Grupo Branco	branquear**; embranquecer**; branquejar; esbranquiçar; alvidecer	4 7	99 2	5 2	9 5	**431** 46	**12** 24
Grupos Castanho, Creme, Prateado e Azul	sem verbos nas formas finitas no CHAVE e CONDIV; pratear; azular; acastanhar, amarronzar	—	—	—	—		

Uma conclusão possível, e que pode ser explorada futuramente com a análise detalhada dos verbos de cores no Vercial, é a perda gradual da propriedade descritiva dos verbos de cores – como se fossem tomando a forma de adjetivos/particípios e convencionalizando-a em direções mais concretas.

3.3 Os usos das cores

Nesta seção relatamos os dados obtidos na exploração dos usos dos verbos de cor. Especificamente, buscamos os argumentos associados aos verbos coloridos, nas posições de sujeito e de objeto.

Com relação ao grupo Dourado, o primeiro dado a chamar a atenção é que quase metade das ocorrências de *dourar/doirar*, em ambas as variantes, não é colorida, mas se refere à expressão "dourar/doirar a pílula", em um uso claramente original.

A tabela 3 apresenta detalhadamente os usos de *dourar/doirar*. Como se pode perceber, são (quase) todos não descritivos: doura-se a *imagem*, a *perspectiva*, o *perdão*, o *brasão*. A única ocorrência que talvez possa ser considerada descritiva, isto é, colorida, diz respeito a "dourar a sala", em ".. o sol doira a sala...". No entanto, consideramos esse (mais) um exemplo de difícil decisão, que pode ser visto tanto do ponto de vista da cor (*o sol deixar a sala da cor dourada*) como um uso não colorido, (*o sol ilumina a sala*). Não há como ter certeza da leitura, e consideramos, portanto, a vagueza uma propriedade inerente e importante da linguagem, que não deve ser eliminada ou vista como imperfeição. Como a língua não nos obriga a escolher uma única interpretação, não o faremos. Especificamente com relação à anotação, estamos diante de um exemplo em que o verbo recebe as duas informações semânticas, simultaneamente: será considerado

tanto [sema="cor"] (para a interpretação colorida), quanto [sema="cor:original"] (para a interpretação não colorida).

Na variante BR, a situação é idêntica, em que se doura, além da pílula, a *imagem*, a *ignorância* e a *personalidade*, entre outros. Assim como na variante PT, o *sol* doura coisas como *paisagem*, *cidade* e *gente*. E, diferente da variante PT, no Brasil também se doura *alho*, *bacon*, *cebola*, etc, isto é, temos o uso associado ao domínio culinário, e, portanto, um uso original – o que não acontece na variante PT, em que o verbo associado a tais ações é *alourar*.[7]

Com relação aos sujeitos de *dourar*, são pouco frequentes, dada a alta ocorrência da forma infinitiva "dourar a pílula". Ainda assim, na parte portuguesa do CHAVE apenas nomes próprios (de pessoas e instituições) exercem a função de sujeito, o que reforça a ideia de um uso original. Na variante BR, além de nome próprio, são *mestres*, *estatutos*, *virtudes*, *retórica* e *jornal* aqueles que *douram*, mais uma vez em um uso original. Ainda na variante BR, *discos de massa*, *carne* e *grill douram*, repetindo-se o uso original vinculado à culinária, e encontramos apenas uma ocorrência de um uso possivelmente colorido, em que a ação de dourar é feita pelo sol, como já mencionado.[8]

[7] Por outro lado, não podemos desconsiderar o fato de haver muito mais notícias sobre comida/culinária no jornal Folha de S. Paulo que no Público – o que não invalida nossa explicação. São apenas 6 os casos de "alourar" na parte PT do CHAVE, e 1 na parte BR (ou 2, se considerarmos também "aloirar").

[8] Para realizar a busca por lema dos sujeitos de dourar ou doirar no AC/DC, após a escolha do corpo desejado para a pesquisa, a expressão de busca é (o exemplo refere-se à busca apenas na variante BR) [func="SUBJ>.*"] [pos!="V.*"]* [pos="V.*" & lema="do[ui]rar" & sema="cor.*" & temcagr!="PCP" & variante="BR"] within s

TABELA 3
Objetos de *dourar/doirar* nos corpos CHAVE e CONDIV[9]

Lema dos objetos de *dourar | doirar*	Distribuição por variante e corpo				Campo semântico
	BR*		PT**		
	CHAVE	CONDIV	CHAVE	CONDIV	
pílula* **	12	—	9	1	original
imagem* ** ; perspectiva**; brasão**; asa**; sala **; chamado**; perdão**; ignorância*; fisiologia*; imagem*; década*; personalidade*; jogada*; triunfo*	5	2	9	—	original
alho*; bacon*; assado*; cebola*; manteiga*; frango*	7	—	—	—	original (culinária)
paisagem*; cidade*; luz*; gente*	3	2	—	—	original | colorido
TOTAL	27	4	18	1	

A análise dos verbos *branquear* e *embranquecer*, pertencentes ao grupo Branco, também revela fatos curiosos. Se voltarmos à tabela 2, vemos que, no CHAVE, tomando-se o lema *branquear*, é enorme a diferença entre a variante BR (4) e a variante PT (99). Quando exploramos essa forma verbal, percebemos que, das 99 ocorrências, 68 referem-se a um uso original da cor, em que o principal objeto de *branquear* é *dinheiro*, e os demais lemas são variados (tabela 4). O segundo uso mais comum de *branquear*, e que também comparece de maneira mais frequente na variante PT, diz respeito ao campo da cor humana. Branqueiam-se *dente*, *perna*, *pele mão*. Ainda na tabela 4, vemos que, na variante BR, não há uso original de *branquear*: o uso colorido é o mais comum, mas ainda assim são poucas as ocorrências no corpos (apenas 3). Com relação ao uso associado ao ser humano, a variante BR conta com apenas uma ocorrência (*branquear o pescoço*).

[9] Para realizar a busca por lema dos objetos de dourar | doirar no AC/DC, após a escolha do corpo desejado para a pesquisa, a expressão de busca é (o exemplo refere-se à busca apenas na variante PT) [sema="cor.*" & grupo="Dourado" & temcagr!="PCP" & pos="V.*" & variante="PT"] [pos!="V.*"]* @[func="<ACC"] within s

Não encontramos nem uma ocorrência de *branquear* usado de maneira original. Com relação à expressão *branquear dinheiro*, muito frequente em PT, o seu equivalente em BR é *lavar dinheiro* (22 ocorrências no CHAVE), ou, mais especificamente, a expressão *lavagem de dinheiro* (96 ocorrências no CHAVE).

TABELA 4
Objetos de *branquear* nos corpos CHAVE e CONDIV[10]

Lema dos objetos de *branquear*	Distribuição por variante e corpo				Campo semântico
	BR*		PT**		
	CHAVE	CONDIV	CHAVE	CONDIV	
dinheiro**	—-	—	14	1	original
imagem**; PIDE**; comportamento**; situação**; fascismo**; derrota**; tortura** etc.	—	—	68	—	original
roupa* **; rolha*; celulose*; marfim**	3	—-	1	1	colorido
dente**; pescoço*; perna**, pele**, mão**	—-	1	1	6	cor:humana
TOTAL	3	1	84	8	

Já a exploração de *embranquecer* aponta para um quadro diferente. Voltando à tabela 2, não apenas a frequência de uso é menor do que em *branquear*, como também os dados são mais equilibrados entre as variantes. A análise qualitativa de *embranquecer* revela um uso associado à cor humana, independente da variante, como pode ser observado na tabela 5.

[10] Expressão de busca utilizada, considerando a variante PT: [sema="cor.*" & lema="branquear" & temcagr!="PCP" & pos="V.*" & variante="PT"] [pos!="V.*"]* @[func="<ACC"] within s
Distribuição de lema

TABELA 5
Objetos de *embranquecer* nos corpos CHAVE e CONDIV

Lema dos objetos de *embranquecer*	Distribuição por variante e corpo				Campo semântico
	BR*		PT**		
	CHAVE	CONDIV	CHAVE	CONDIV	
tudo e todos	1	—	—	—	cor:raça | original
Pele* **; mão**; cabelo* **	2	—	1	2	cor:humana
TOTAL	3	—	1	2	

A exceção é para uma ocorrência de *embranquecer*, na variante BR, que pode ser associada a um uso original ou, simultaneamente, ao uso associado à utilização da cor como raça, como pode ser percebido na frase:

> FSP951120-009: É que foram antes massacrados por uma mídia que busca embranquecer **tudo e todos**.

Por fim, se desconsideramos, na busca, a presença de objeto, e analisamos formas intransitivas de *embranquecer*, aparecem mais duas ocorrências, ambas na variante BR, nenhuma colorida, e ambas associadas ao campo semântico da raça:

> FSP950319-128: Ao adotar como visão de si mesmo a ideologia de seus dominadores, o mestiço opta pelo recalque e pela traição de tudo que nele não for espelho da Europa: o filho de índia e europeu, identificando-se com o pai, tornou-se perseguidor do gentio materno; o mulato, buscando ascender socialmente, trata desesperadamente de **embranquecer**, reforçando e legitimando o preconceito com o negro.
> FSP950623-092: Suassuna acredita que os espetáculos, se bem produzidos, seriam capazes de atrair ' 300 vezes mais gente do que o Michael Jackson» (que ele define como ' traidor da raça negra», por sua suposta vontade de ' **embranquecer**») .

Com relação ao grupo Preto, nossa análise considerou apenas o verbo *enegrecer*. Pela tabela 2, constatamos que *enegrecer* é muito mais frequente na variante PT (e com um uso muitíssimo mais alto no corpo Vercial (63), que não exploramos aqui), e é usado principalmente de maneira original. A tabela 6 oferece uma visão qualitativa de *enegrecer*, em ambas as variantes.

TABELA 6
Objetos de enegrecer nos corpos CHAVE e CONDIV

Lema dos objetos de *enegrecer*	Distribuição por variante e corpo				Campo semântico
	BR*		PT**		
	CHAVE	CONDIV	CHAVE	CONDIV	
costa**, céu**, prata**	—	—	2	1	cor
Abril**; coração**; imagem**; verso**; cenário* **; desporto**	1	—	6	1	cor:original
dente**	—	—	1	—	cor:humana
TOTAL	1	—	9	2	

Quando buscamos pelos responsáveis pelo enegrecimento, encontramos, nos usos coloridos, *petróleo* (*enegrece a costa*), *fumo* (*enegrece o céu*) e *enxofre* (*enegrece a prata*) e, nos usos originais, *ódio* (*enegrece o coração*), ou incertezas (*enegrecem o cenário*), sendo apenas o último exemplo relativo à variante BR. Se eliminamos a restrição referente à presença de objeto, temos mais uma ocorrência de *enegrecer*, na variante BR do CHAVE, vaga entre [sema="cor:raça"] ou [sema="cor:humana"]:

> FSP951016-045: No começo do século, sociólogos acreditavam fortemente nas variáveis raciais e, otimistas, diziam que o Brasil não tinha um problema com os negros, pois iríamos branqueando gradativamente com os casamentos inter-raciais, enquanto os americanos se preocupavam, pois acreditavam que iriam **enegrecendo** por meio dos mesmos cruzamentos entre negros e brancos.

Os verbos do grupo Amarelo – *amarelar* e *amarelecer* – também apresentam comportamentos distintos conforme a variante.

Com relação a *amarelar*, em termos quantitativos, quase não há diferença entre as variantes PT e BR (tabela 2). No entanto, a análise qualitativa a partir dos corpos (tabela 7) revela que, considerando apenas o corpo CHAVE-BR encontramos (poucas) ocorrências transitivas[11] de *amarelar*, que é usado sobretudo relacionado a atributos humanos, mas encontramos também um uso original (*o remorso amarelando os olhos*) e um uso colorido.

No entanto, se consideramos os usos intransitivos ou transitivos indiretos de *amarelar*, o quadro se altera ligeiramente.

TABELA 7
Objetos de amarelar nos corpos CHAVE e CONDIV

Lema dos objetos de *amarelar*	Distribuição por variante e corpo				Campo semântico
	BR*		PT**		
	CHAVE	CONDIV	CHAVE	CONDIV	
Pratos*	1	—	—	—	cor
Pessoas*; unhas*	2	—	—	—	cor:humana
olhos*	1				cor: original
TOTAL	4	—	—	—	

Encontramos mais três ocorrências (duas no CHAVE BR, uma no CONDIV PT), todas em um uso claramente original, mas com diferenças de sentido conforme a variante. Na variante BR, *amarelar* equivale a *ter medo de*; *se acovardar* (sentido inexistente na variante PT). Na variante PT (cf. o terceiro exemplo), *amarelar* equivale a *distribuir cartões amarelos*.

> FSP951030-086: Folha — Você já «**amarelou**» para ondas grandes em alguma competição? (CHAVE-BR)
> par=fut48437: Mas não é menos que, depois que descobriram que a capital boliviana era nos Andes, os jogadores brasileiros literalmente **amarelaram** .(CONDIV-BR)

[11] [sema="cor.*" & lema="amarelar" & temcagr!="PCP" & pos="V.*" & variante="PT"] [pos!="V.*"]* @[func="<ACC"] within s

par=fut18426: O jogo foi muito árduo e houve necessidade de «**amarelar**». (CONDIV- PT)

Com relação a *amarelecer*, a tabela 2 já indica o uso mais frequente na variante PT, com apenas 3 ocorrências na variante BR. A tabela 8 contém a descrição qualitativa do comportamento de *amarelecer*.

TABELA 8
Objetos de amarelecer nos corpos CHAVE e CONDIV

Lema dos objetos de *amarelecer*	Distribuição por variante e corpo				Campo semântico
	BR*		PT**		
	CHAVE	CONDIV	CHAVE	CONDIV	
espetáculo*; mármore**	1	—	2	—	cor
pele*	1	—	—	—	cor:humana
TOTAL	2	—	—	—	

No entanto, como *amarelecer* é usado principalmente de maneira intransitiva, apresentamos os sujeitos de *amarelecer*, na variante PT: *página*, *tinta*, *relva*, *colectâneas*, *relvado* [sema="cor"]; *fruta*, *sorriso* [sema="cor:original"]. Notamos ainda uma ocorrência de *amarelecer de inveja*, referente a [sema="cor:original"].

Por fim, detalhamos os usos dos verbos do grupo Verde – *esverdear*, *reverdecer e verdejar*. A tabela 2 mostra que, para os três lemas, há muito mais ocorrências na variante PT. No entanto, se restringimos a análise aos corpos CHAVE e BR, as diferenças somem, com no máximo duas ocorrências por variante. A única diferença está no lema *esverdear*, na variante PT do CHAVE, com 8 ocorrências, que detalhamos a seguir. Todas as ocorrências de *esverdear* são originais, como vemos na tabela 9:

TABELA 9
Objetos de esverdear nos corpos CHAVE e CONDIV

Lema dos objetos de *esverdear*	Distribuição por variante e corpo				Campo semântico
	BR*		PT**		
	CHAVE	CONDIV	CHAVE	CONDIV	
indústria**; pílula**; sistema**; projetos**; atividade econômica**	—	—	7	1	Cor:original

Com relação aos demais verbos de cor – *rosar*, *acinzentar*, *arroxear*; *avermelhar*; *alaranjar* – a baixa frequência de ocorrência nos corpos CHAVE e CONDIV e o limite de páginas nos obrigam a postergar a descrição para outra ocasião. Lembramos, no entanto, que todo o material encontra-se publicamente disponível na página do projeto AC/DC para aqueles que se interessarem pelo campo das cores.

4 Considerações finais

Neste trabalho, exploramos dois corpos jornalísticos, com textos das variantes de Portugal e do Brasil, quanto a classes gramaticais que se referem às cores, com especial atenção aos verbos.

De maneira sumarizada, os resultados encontrados indicam que:

- A distribuição das categorias gramaticais é a mesma entre as variantes portuguesa e brasileira;
- A distribuição dos verbos de cor e dos adjetivos de cor também é a mesma entre as variantes portuguesa e brasileira;
- Como esperado, verbos se prestam pouco a representar o processo de "colorização" das coisas. Assim, não é surpresa que a grande maioria dos verbos de cores esteja no particípio. E, quando desconsideramos essa forma, independente de variante ou gênero, certos grupos de cores, como *Castanho* e *Creme* desaparecem. Por outro lado, grupos, como *Verde*, *Branco*, *Preto*, *Amarelo* e *Dourado*, são mais produtivos na formação de verbos, dando origem a dois ou mais lemas distintos.

Especificamente quanto às formas verbais que se referem às cores, acreditamos ter descoberto pontos que merecem análise mais detalhada. De uma perspectiva do uso, é interessante perceber o que fazem os verbos de cor. *Branquear* e *embranquecer*, por exemplo, têm comportamentos diferentes não apenas quanto à frequência. *Branquear*, mais comum, é usado principalmente na variante de Portugal, e de maneira metafórica: muito mais que os *dente*s, branqueiam-se *dinheiro* e *imagem*. *Embranquecer* tem um uso mais literal – embranquecemos os *cabelos*, a *pele*. No entanto, o que mais nos chama a atenção é a frequência dos usos originais associados aos verbos de cor – excluída a forma participial –, o que parece sugerir uma perda gradativa da propriedade descritiva desses verbos. A alta ocorrência, em muitos casos, dos verbos de cor apenas no corpo VERCIAL, composto por textos literários antigos, reforça essa hipótese.

Alguns pontos permanecem para análise futura: nossa busca excluiu locuções verbais. Usamos mais "ficar vermelho" do que "avermelhar"? E o que nos diria a análise das cores no particípio?

Por fim, reforçamos a ideia de que para comparar termos de cor entre duas ou mais línguas é importante uma caracterização detalhada das cores em cada uma das línguas contrastadas. Em nosso estudo, buscamos conjugar as dimensões de uso e de forma, oferecendo um quadro das cores em português, considerando também distinções entre as variantes do Brasil e de Portugal.

Agradecimentos

O presente trabalho foi desenvolvido no âmbito da Linguateca, financiada pelo governo português, UMIC, FCCN e União Europeia (FEDER e FSE), por meio de POSC/339/1.3/C/NAC, e pela FCT.

Referências

BACELAR DO NASCIMENTO, Maria Fernanda; CARVALHO, Anabela . Preto e branco ou branco e preto? (Como se combinam os nomes de cores). XI ENCONTRO NACIONAL DA ASSOCIAÇÃO PORTUGUESA DE LINGUÍSTICA, *Actas...* Lisboa, 2-4 de Outubro de 1995, 1996, Lisboa: APL/Colibri, p. 367-380.

BERLIN, Brent; KAY, Paul. *Basic Colour Terms:* their Universality and Evolution. Stanford: CSLI, 1991 [1. edição: 1969]

BIDERMAN, Maria Tereza Camargo; NASCIMENTO, Maria Fernanda Bacelar do ; PEREIRA, Luisa Alice Santos. Uso das cores no português brasileiro e no português europeu. In: ISQUIERDO, Aparecida Negri; ALVES, Ieda Maria (Ed.). *As ciências do léxico:* Lexicologia, lexicografia, terminologia, vol. III, Editora UFMS, Associação editorial Humanitas, 2007. p. 105-124.

CORREIA, Margarita. Towards a General Description of the Semantic Field of 'Colour' in European Portuguese. In: BIGGAM, C. P.; KAY, Christian J. (Ed.). *Progress in Colour Studies*, 1: Language and Culture. Amsterdam/Filadélfia: John Benjamins, 2006. p. 111-25.

COSTA, Luís, Diana Santos; ROCHA, Paulo Alexandre. Estudando o português tal como é usado: o serviço AC/DC. THE 7TH BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL 2009), *Anais...* 2009, São Carlos, Brasil.

DEUTSCHER, Guy. *Through the language glass*: Why the World Looks Different in Other Languages. Metropolitan: Henry Holt, 2010.

FARIAS, Emília Maria Peixoto; MARCUSCHI, Luís Antônio. A metáfora das cores na linguagem e no pensamento. In: PINTO, Abuêndia Padilha. (Org.). *Tópicos em cognição e linguagem*. 1. ed. Recife: Ed. Universitária da UFPE, 2006, v. 1. p. 19-55.

FRANKENBERG-GARCIA, Ana; SANTOS, Diana. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução* IX.1, p. 61-79, 2002. Universidade de Santa Catarina. ISSN: 1676-7047.

INÁCIO, Susana; SANTOS, Diana; SILVA, Rosário. COMPARAndo cores em português e inglês. In: FROTA, Sónia; SANTOS, Ana Lúcia. (Org.). *Artigos selecionados do XXIII Encontro da Associação Portuguesa de Linguística*, APL, 2008. p. 271-86.

JORGE, Guilhermina (Coord.). et al. As cores preto no branco: uma análise comparativa. *Polifonia* 6, Revista do Grupo Universitário de Investigação em Línguas Vivas da Universidade de Lisboa, Edições Colibri, Lisboa, p. 119-133, 2003.

LUCY, John A. Linguistic relativity. *Annual Review of Anthropology* 26. Palo Alto, CA: Annual Reviews Inc, p. 291-312, 1997.

MOTA, Cristina; SANTOS, Diana. *Corte e costura no AC/DC*: auxiliando a melhoria da anotação nos corpos. Setembro de 2009. Disponível em: http://www.linguateca.pt/acesso/corte-e-costura.pdf.

PHILIP, Gillian Susan. *Collocation and connotation*: A Corpus-Based Investigation Of Colour Words In English And Italian. PhD thesis, University of Birmingham, Março de 2003.

PHILIP, Gill. *Colouring Meaning*: Collocation and connotation in figurative language. John Benjamins Publishing Company, 2011.

ROCHA, Paulo; SANTOS, Diana. CLEF: Abrindo a porta à participação internacional em avaliação de RI do português. In: SANTOS, Diana. (Ed.). *Avaliação conjunta*: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal: IST Press, 2007. p. 143-158.

ROSCH, Eleanor. Universals and cultural specifics in human categorization. In: BRISLIN, Richard; BOCHNER, Stephen; LONNER, Walter. (Ed.). *Cross-cultural perspectives on learning*. Wiley, 1975. p. 177-206.

SANTOS, Diana. Corpora at Linguateca: vision and roads taken. In: SARDINHA, Tony Berber; FERREIRA, Telma São Bento (Ed.). *Working with Portuguese corpora*. No prelo.

SANTOS, Diana; MOTA, Cristina. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. In: CALZOLARI, Nicoletta; CHOUKRI, Khalid; MAEGAARD, Bente; MARIANI, Joseph; ODIJK, Jan; PIPERIDIS, Stelios; ROSNER, Mike; TAPIAS, Daniel. (Ed.). *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, 17-23 de Maio de 2010), European Language Resources Association, p. 1437-1444.

SANTOS, Diana; SILVA, Rosário; INÁCIO, Susana. What's in a colour? Studying and contrasting colours with COMPARA. SIXTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2008). *Proceedings...* (Marraqueche, 26 de Maio -1 de Junho de 2008), ELDA.

SANTOS, Diana; SILVA, Rosário; FREITAS, Cláudia. Pluralidades na cor: contrastando a língua do Brasil e de Portugal. In: SILVA, Augusto Soares da; TORRES, Amadeu; GONÇALVES, Miguel. (Org.). *Línguas Pluricêntricas*: Variação Linguística e Dimensões Sociocognitivas. Pluricentric Languages: Linguistic Variation and Sociocognitive Dimensions. Braga: Aletheia, Publicações da Faculdade de Filosofia da Universidade Católica Portuguesa, 2011. p. 555-572.

SANTOS, Diana, Stella E.O. Tagnin & Elisa Duarte Teixeira. "Colours, clothing and food in CorTrad: why corpus-based translation studies are revealing". Apresentação no ICAME 2011 (Oslo, 1-5 de junho de 2011).

SAUNDERS, Barbara. Revisiting basic colour terms. *Science as Culture*, *The Human Nature Review*, 2005. Disponível em: http://human-nature.com/science-as-culture/saunders.html.

SILVA, Augusto Soares da. O corpus CONDIV e o estudo da convergência e divergência entre variedades do português. In: COSTA, Luís; SANTOS, Diana; CARDOSO, Nuno (Ed.). *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008. p. 25-28.

SILVA, Rosário; SANTOS, Diana. *Arco-íris*: notas sobre a anotação do campo semântico da cor em português. Primeira edição: 25 de junho de 2009. Versão atual: 24 de janeiro de 2012. Disponível em: http://www.linguateca.pt/acesso/ArcoIris.pdf.

SILVA, Rosário; INÁCIO, Susana; SANTOS, Diana. *Documentação da anotação relativa à cor no COMPARA*. Última versão: 31 de Dezembro de 2008. Disponível em: http://www.linguateca.pt/COMPARA/DocAnotacaoCorCOMPARA.pdf

SLETSJØE, Leif. Sobre el topico de los ojos verdes. *Strenae*: Estudios de filologia e historia dedicados al profesor Manuel Garcia Blanco. Salamanca, 1962. p. 445-459.

SWEARINGEN, Andrew. *Seeing red in Roxo*: The Evolution of Portuguese colour terms. MA thesis, Univ. of Copenhagen, 2000.

TEIXEIRA, Elisa D.; SANTOS, Diana; TAGNIN, Stella E. O. CorTrad: um novo corpus paralelo multiversão para o par de línguas português-inglês. In: SHEPHERD, Tania M.G.; SARDINHA, Tony Berber; PINTO, Marcia Veirano. (Ed.). *Caminhos na Linguística de Corpus*. Mercado de Letras, 2012. p. 151-176.

WIERZBICKA, Anna. The meaning of color terms: semantics, culture, and cognition. *Cognitive Linguistics*, 1-1, p. 99-150, 1990.

ZAVAGLIA, Cláudia. Dicionário e cores. *Alfa* 50, 2, São Paulo, p. 25-41, 2006.

ZAVAGLIA, Cláudia. A prática lexicográfica multilingue: questões concernentes ao campo das cores. In: ISQUIERDO, Aparecida Negri; ALVES, Ieda Maria. (Ed.). *As ciências do léxico*: Lexicologia, lexicografia, terminologia, vol. III, Editora UFMS, Associação Editorial Humanitas, 2007. p. 209-222.

III FRAMES, CONSTRUÇÕES E ANOTAÇÃO DE CORPORA

Desafios para a anotação semântica de textos jurídicos: limites no uso da FrameNet e rotas alternativas

Anderson Bertoldi[1]
Rove Chishman[2]

RESUMO: Este trabalho objetiva analisar a aplicabilidade das etiquetas semânticas da FrameNet para anotação de *corpora* jurídicos, considerando-se a variante brasileira do português e o sistema judiciário brasileiro. Para alcançar esse objetivo, as unidades lexicais evocadoras do *frame* Criminal_process são contrastadas com seus equivalentes em português e discutidas as dificuldades encontradas no uso das etiquetas semânticas da FrameNet. No caso de anotação de *corpora* jurídicos, a falta de correspondência entre os sistemas jurídicos norte-americano e brasileiro traz à tona a necessidade de se criar um recurso lexical baseado em *frames* para a linguagem jurídica brasileira que possa ser utilizado para a anotação de *corpora*.

PALAVRAS-CHAVE: Semântica de *Frames*, anotação semântica, recursos lexicais jurídicos.

ABSTRACT: The objective of this paper is to evalute the applicability of Frame Semantics for the annotation of legal corpora, taking into accont the Brazilian Portuguese and the Brazilian legal system. The lexical units evoquing Criminal_process frame are contrasted with the Portuguese equivalents. The difficulties found in the annotation of those sentences using the FrameNet semantic tags are discussed in this paper. Considering the annotation of legal corpora, the lack of correspondence between the American and the Brazilian Legal systems brought out the necessity of developing a frame-based lexical

[1] Anderson Bertoldi é doutor em Linguística Aplicada pela Unisinos. No momento, realiza pós-doutorado junto ao PPG de Linguística Aplicada da Unisinos. E-mail para contato: andersonbertoldi@yahoo.com.

[2] Rove Chishman é doutora em Letras pela PUCRS e atua como professora do PPG em Linguística Aplicada da Unisinos. E-mail para contato: rove@unisinos.br.

resource for the Brazilian legal language. Such a lexical resource could be used for legal corpora annotation of Brazilian legal texts.

KEYWORDS: Frame Semantics, corpora annotation, legal lexical resourses.

1 Introdução

O objetivo deste artigo é avaliar a aplicabilidade da Semântica de *Frames* e do paradigma FrameNet para a anotação de textos jurídicos.[3] A Semântica de *Frames* (FILLMORE, 1982, 1985) é uma teoria da linguística cognitiva que busca explicar o complexo sistema de relações cognitivas que o falante precisa saber a fim de entender um enunciado. Entender que algo foi comprado requer o entendimento de que algo foi vendido. Assim, entender o cenário de uma transação comercial requer a identificação dos diferentes participantes de um evento comercial, tais como o *comprador*, o *vendedor* e o *produto*. A Semântica de *Frames* não faz distinção entre conhecimento linguístico e conhecimento de mundo. Em outras palavras, o conhecimento do significado das palavras *comprar* e *vender* implica o conhecimento cultural relacionado a um evento comercial: o comprador necessita de dinheiro para comprar um produto e o vendedor espera receber dinheiro por um produto vendido.

A FrameNet é uma base de dados lexicais que descreve o significado das palavras relacionando-as a um *frame* semântico. As unidades primárias de descrição lexical na FrameNet são o *frame* e a unidade lexical (FILLMORE; JOHNSON; PETRUCK, 2003). O *frame* é um sistema de conceitos relacionados de tal forma que, para se entender um conceito, é necessária a compreensão de todos os conceitos relacionados (FILLMORE, 1982). A unidade lexical é compreendida como o emparelhamento de uma palavra com um *frame* semântico. Na FrameNet, a informação sobre valência é especificada em dois níveis: o sintático e o semântico. A valência sintática especifica os tipos sintagmáticos (sintagma nominal, preposicional etc.) e as funções gramaticais (sujeito, objeto etc.). A

[3] Este trabalho foi desenvolvido no âmbito do projeto *Tecnologias Semânticas e Sistemas de Recuperação de Informação Jurídica*, financiado pela CAPES através do Edital Nº. 020/2010/CAPES/CNJ. Este trabalho também contou com o financiamento das agências CAPES, CNPq e FINEP, através do Edital Nº. 001/2010, MCT/CNPq/FINEP – Programa Nacional de Pós-Doutorado (PNPD).

valência semântica é descrita em termos de entidades que podem participar de um *frame* evocado por uma unidade lexical, tais entidades são chamadas de "elementos de *frame*" (FILLMORE; JOHNSON; PETRUCK, 2003).

Diferentes projetos têm utilizado a FrameNet para anotação de textos em diferentes línguas. Apesar de as etiquetas semânticas da FrameNet serem utilizadas para a anotação de textos em diferentes línguas, essas etiquetas foram criadas a partir da análise de unidades lexicais em inglês. O projeto SALSA (*Saarbrücken Lexical Semantics Annotation and Analysis*) vem realizando a anotação de *corpora* em língua alemã a partir das etiquetas semânticas da FrameNet (BURCHARDT et al., 2009). O projeto SALSA parte da suposição de que é possível reutilizar as etiquetas semânticas da FrameNet para a análise semântica do alemão. Assim, esse projeto inclui (a) a anotação de um *corpus* de grande porte em alemão e a geração de léxico baseado em *frames* a partir da anotação do *corpus* e (b) a indução de modelos baseados em dados para análise semântica automática e aplicações em processamento de linguagem natural (BURCHARDT et al., 2009).

O trabalho do projeto SALSA com a anotação de *corpora* de grande extensão, geração automática de entradas lexicais a partir de *corpora* anotados e análise semântica automática influenciou vários trabalhos de criação automática de FrameNet. A proposta de Padó e Lapata (2005) sugere o uso de *corpora* paralelos para a criação automática de entradas lexicais baseadas em *frames*. A partir da anotação de um *corpus* em inglês com as etiquetas da FrameNet, seria possível transferir a anotação do *corpus* em inglês para um *corpus* de outra língua. Essa técnica vai inspirar vários trabalhos de transferência automática de anotação semântica, como é o caso dos trabalhos nas línguas francesa (PADÓ; PITEL, 2007) e italiana (TONELLI; PIANTA, 2008; DINI; BOSCA, 2009; VENTURI et al., 2009).

Este trabalho objetiva mostra como as diferenças entre sistemas jurídicos diferentes podem ocasionar algumas divergências no que tange à aplicação das etiquetas da FrameNet para anotação de textos jurídicos em português do Brasil. Para alcançar este objetivo, as unidades lexicais evocadoras do *frame* Criminal_process[4] foram contrastadas com seus equivalentes em português. As unidades

[4] Seguindo a padronização sugerida pelos pesquisadores da FrameNet, neste trabalho os nomes dos *frames* são escritos em fonte Courier New e os nomes dos elementos de *frame* são escritos em caixa alta.

lexicais evocadoras de onze *subframes* que compõem o *frame* Criminal_process foram contrastadas com seus equivalentes em português, a fim de se verificar se o contexto jurídico (o *frame*) evocado pela unidade lexical em inglês era semelhante ao contexto jurídico evocado pela unidade lexical em português.

Assim, para discutir o uso da Semântica de *Frames* e do paradigma FrameNet para a anotação de textos jurídicos em português brasileiro, o presente artigo está estruturado conforme segue. Na seção 2 são apresentados os princípios da Semântica de *Frames*. Na seção 3 é apresentada a base de dados lexicais FrameNet. Na seção 4 são discutidos os limites para o uso das etiquetas semânticas da FrameNet para anotação de textos jurídicos. A seção 5 apresenta as direções futuros deste trabalho.

2 A Semântica de *Frames*

A Semântica de *Frames* nasce a partir de um conceito muito discutido na década de 70, o ***frame*** (MINSKY, 1974; GOFFMAN, 1974). Inicialmente, Fillmore (1975) faz uma distinção entre os conceitos de **cena** e ***frame***. A **cena** seria não apenas uma cena visual, mas todo um conjunto de tipos familiares de transações interpessoais, cenários padrões definidos culturalmente, estruturas institucionais, experiências inativas, imagem corporal, crenças humanas, ações experiências e imagens. O ***frame*** seria um sistema de escolhas linguísticas, sejam palavras, regras ou categorias gramaticais, associadas à determinada instância prototípica de uma cena.

Fillmore (1977) demonstra, através do evento de transação comercial, que os verbos *comprar*, *vender* e *custar* representam diferentes perspectivas do mesmo evento. O vendedor cede a mercadoria em troca de dinheiro e o comprador cede o dinheiro em troca da mercadoria. Um evento como transação comercial marca a troca de posse de dois bens: o dinheiro passa da posse do comprador para o vendedor e a mercadoria passa da posse do vendedor para o comprador. A análise da relação de perspectiva no evento de transação comercial já apresenta um primeiro esboço do que será chamado de **elementos de *frame*** posteriormente. Esses elementos de *frame* vêm substituir a proposta de casos (FILLMORE, 1968).

A distinção entre cena como estrutura cognitiva e *frame* como estrutura linguística é posteriormente abandonada (FILLMORE, 1982,

1985). Segundo Fillmore (1982, p.111), "pelo termo '*frame*' eu tenho em mente qualquer sistema de conceitos relacionados de tal forma que para entender qualquer um deles você tem que entender toda a estrutura na qual ele se encaixa (...)". Para a Semântica de *Frames*, as palavras têm a capacidade de "evocar" todo um conhecimento de mundo que é organizado através de uma estrutura cognitiva chamada de *frame*: "Um *frame* é evocado pelo texto se alguma forma ou padrão linguístico é convencionalmente associado com o *frame* em questão (FILLMORE, 1985, p.232)".

Fillmore e Atkins (1992) apresentam o primeiro exercício de análise semântica baseada em *frames* e apontam a futura criação de um dicionário *on-line* baseado em *frames*. A partir do estudo de unidades lexicais que expressam risco, como *risk*, *danger* e *hazard*, Fillmore e Atkins propõem onze categorias para descrever os participantes do *frame Risk*. Essas categorias são: *chance*, *harm*, *victim*, *valued object*, *risky situation*, *deed*, *actor*, *intended gain*, *purpose*, *beneficiary* e *motivation*. No entanto, esses papéis semânticos ainda não são chamados de elementos de *frame*, nome que será dado, posteriormente, aos papéis semânticos desenvolvidos no contexto do Projeto FrameNet.

3 A FrameNet e os *frames* semânticos

A FrameNet é um projeto lexicográfico desenvolvido pelo *International Computer Science Institute*, Berkeley, desde 1998. O nome FrameNet é inspirado na WordNet (MILLER, 1995; FELLBAUM, 1998), um recurso lexical que organiza as palavras segundo relações semânticas. A FrameNet é um recurso lexical que descreve o significado segundo os princípios da Semântica de *Frames*. Conforme a Semântica de *Frames*, as formas lingüísticas evocam a informação contextual "armazenada" na estrutura cognitiva conhecida como *frame*.

Os itens lexicais na FrameNet são tratados como unidades lexicais. A unidade lexical é vista como o emparelhamento de uma palavras com um *frame* semântico, não a palavra em si. Cada novo significado de uma palavra representa uma nova unidade lexical. Assim, são as unidades lexicais que evocam os *frames*. Embora a FrameNet não trate da polissemia, segundo a Semântica de *Frames*, a polissemia é pode ser concebida como uma palavra que apresenta diferentes unidades lexicais.

O método de analise lexical adotado pela FrameNet, conforme Fillmore e Baker (2010), segue cinco etapas:

1. **Caracterização do *frame***. Caracteriza-se a situação descrita pelas unidades lexicais, por exemplo, a prisão de um suspeito, como no caso do *frame* Arrest.

   A acusa B de ter cometido um crime e A prende B.

2. **Descrição e nomeação dos elementos de *frame***. Após a caracterização de um *frame* específico, identificam-se todos os possíveis participantes da situação e criam-se nomes para cada participante.

   No caso do *frame* Arrest, A é chamado de AUTORIDADES, B é chamado de SUSPEITO, o crime é chamado de OFENSA e a acusação contra o suspeito é chamada de ACUSAÇÕES.

3. **Seleção das unidades lexicais**. Após a descrição da situação e da identificação e nomeação dos elementos de *frame*, as unidades lexicais e expressões evocadoras do *frame* são identificadas. No caso do *frame* Arrest, as unidades lexicais elencadas pela FrameNet são:

   apprehend.v, apprehension.n, arrest.n, arrest.v, book.v, bust.n, bust.v, collar.v, cop.v, nab.v, summons.v

4. **Anotação de sentenças**. Sentenças selecionadas para exemplificar os padrões sintáticos e semânticos de cada unidade lexical são anotadas com elementos de *frame*. As sentenças abaixo demonstram dois exemplos anotados pela FrameNet:

   Are [you $_{\text{AUTORIDADES}}$] **arresting** [me $_{\text{SUSPEITO}}$] [for the murder of Topaz Brown? $_{\text{OFENSA}}$]

   When he gave them his name [they $_{\text{AUTORIDADES}}$] **arrested** [him $_{\text{SUSPEITO}}$] [on a charge of rape. $_{\text{ACUSAÇÕES}}$]

5. **Geração automática de entradas lexicais**. Os exemplos anotados para cada unidade lexical são transformados automaticamente em uma entrada lexical contendo a definição da unidade lexical, as realizações sintáticas de cada elemento de *frame* e os padrões valências.

Na FrameNet, a informação sobre valência é especificada em dois níveis: o sintático e o semântico. A valência sintática especifica os tipos sintagmáticos (sintagma nominal, preposicional etc.) e as funções gramaticais (sujeito, objeto etc.). A valência semântica é descrita em termos de entidades que podem participar de um *frame* evocado por uma palavra, ou seja, os elementos de *frame* (FILLMORE; JOHNSON; PETRUCK, 2003).

Conforme Fillmore e Baker (2010), os elementos de *frame* representam propriedades ou entidades que podem ou devem estar presentes em qualquer instância de um *frame*. A FrameNet diferencia os elementos de *frame* em **centrais**, **periféricos** e **extratemáticos**. Segundo Fillmore e Baker (2010), a distinção entre esses tipos nem sempre é clara. De uma forma geral, elementos de *frame* que são obrigatoriamente expressos são centrais. No caso de verbos, elementos de *frame* que expressam funções sintáticas centrais como sujeito e objeto também devem ser centrais. Em alguns casos, determinados elementos de *frame* que são centrais não necessitam ser expressos. O *frame* `Arrest`, por exemplo, possui como elementos de *frame* centrais AUTORIDADES, SUSPEITO, ACUSAÇÕES e OFENSA. No entanto, quando o elemento ACUSAÇÃO é expresso, o elemento OFENSA é suprimido.

Os elementos de *frame* periféricos expressam em geral funções de adjuntos, expressando tempo, lugar ou modo. Se algumas unidades lexicais expressam lugar como um elemento periférico, outras vão ter o elemento de *frame* indicando lugar como elemento de *frame* central. A diferença entre elementos centrais e periféricos depende essencialmente da necessidade de complementação da unidade lexical. Segundo Fillmore (2007), a diferença entre central e periférico é análoga à distinção entre **actantes** e **circunstantes** apresentada por Tesnière (1959). Os elementos de *frame* extratemáticos introduzem informação referente a outro *frame*, tal como o propósito motivador de algum evento ou ação. Os elementos de *frame* periféricos e extratemáticos são agrupados na FrameNet sob a denominação de elementos não-centrais.

Às vezes, os elementos de *frame* de uma unidade lexical podem ser simplesmente omitidos. Segundo Fillmore e Baker (2010), há uma explicação gramatical e duas lexicais para os elementos de *frame* não realizados. A explicação gramatical está relacionada com estruturas gramaticais que permitem a omissão de algum argumento. Esse tipo de omissão de argumento é chamado de **instanciação nula construcional** (INC). Exemplos de instanciação nula construcional são a omissão do sujeito em orações imperativas e a omissão do agente em oração em voz passiva.

Os outros dois casos são chamados de **instanciação nula indefinida** (INI) e **instanciação nula definida** (IND). Os casos de instanciação nula indefinida envolvem verbos que podem assumir tanto uma forma transitiva como intransitiva, como o verbo *to eat*, em inglês. Nesses casos, a FrameNet trata o verbo como transitivo e considera o objeto omitido como um caso de instanciação nula indefinida. Já os casos de instanciação nula definida envolvem os casos em que o objeto do verbo é essencial, porém ambas as partes envolvidas na conversa o omitem por compartilharem do mesmo conhecimento.

4 Anotação de *corpora* jurídicos

Esta seção aborda os desafios a serem enfrentados ao se usarem as etiquetas semânticas da FrameNet para anotação de *corpora* jurídicos. Apesar de alguns trabalhos afirmarem que os *frames* semânticos podem ser utilizados com sucesso para anotação automática de *corpora* (PADÓ; LAPATA, 2005; PADÓ, 2007), este trabalho explora os casos potenciais de falta de correspondência entre *frames* em inglês e português. Esses casos mostram a necessidade de adaptação dos *frames* semânticos para seu uso em anotação de *corpora* jurídicos. As observações feitas neste trabalho levam em conta a anotação manual, porém, são úteis também para a anotação automática de *corpus* com as etiquetas semânticas da FrameNet.

4.1 Equivalência lexical

O uso das etiquetas da FrameNet para anotação semântica em diferentes línguas requer a reflexão sobre a equivalência das unidades lexicais. Os *corpora* jurídicos apresentam dois pontos de dificuldade:

(a) a equivalência das unidades lexicais e (b) a equivalência dos concecitos jurídicos. Após identificar a unidade lexcial a ser anotada, é necessário encontrar um equivalente em inglês para a unidade lexical que se quer anotar em português. Nessa etapa da anotação, o anotador pode fazer uso do seu conhecimento das línguas em questão ou de um dicionário bilíngue. Uma vez encontrado um equivalente em inglês, é possível pesquisar a base de dados da FrameNet e verificar qual *frame* é evocado pela unidade lexical em inglês.

A equivalência das unidades lexicais é um dos pontos críticos para anotação de *corpora* jurídicos, porque os evento jurídicos no Brasil e nos Estados Unidos nem sempre são correspondentes. Considerando a anotação da unidade lexical *acusar*, a etapa inicial, que consiste em buscar um equivalente em inglês, parece simples: tem-se a unidade correspondente *to accuse*. A etapa seguinte, que consiste em identificar na base de dados *online* da FrameNet os *frames* evocados pela unidade lexical *to accuse*, também não parece problemática. Três *frames* distintos são evocados pelo verbo em questão: `Judgment_communication`, `Judgment` e `Notification_of_charges`. Contudo, observa-se que apenas o *frame* `Notification_of_charges` é relacionado ao domínio jurídico. Este exemplo, ainda que ilustre um caso de anotação que requer do anotador a escolha dos *frames* que dizem respeito ao domínio estudado, trata-se de um caso simples. A próxima etapa envolveria a conferência dos elementos de *frame*. O desafio consiste em verificar se as categorias semânticas que descrevem os eventos jurídicos norte-americanos se prestam para a descrição dos eventos jurídicos brasileiros (a correspondência de *frames* será discutida na seção 4.2 e 4.3).

A unidade lexical *denunciar* representa um caso mais complexo de anotação. O anotador pode se sentir em dúvida ao procurar um equivalente apropriado em inglês. O equivalente mais óbvio para a unidade lexical *denunciar* seria *to denounce*. No entanto, se o anotador procurar pela unidade lexical *to denounce* na FrameNet, ele verá o único *frame* evocado por essa unidade lexical é o *frame* `Judgment_communication`. Como o *frame* `Judgment_communication` não é um *frame* que descreve um evento jurídico, relacionado ao domínio jurídico, o anotador terá que encontrar um outro equivalente para ao unidade lexical *denunciar* e fazer uma nova busca na base de dados da FrameNet. Nesse caso, o anotador pode fazer uso de um dicionário jurídico bilíngue. O resultado possível de uma busca em um dicionário jurídico bilíngue

pode ser uma entrada lexical como na figura 1, extraída do *Dicionário Jurídico*, de Goyos Jr. (1992):

Português	**Inglês**
____ *denunciar*	*to denounce; accuse; inform against; report; proclaim.*

FIGURA 1 - Equivalentes de *Denunciar* (GOYOS Jr., 1992)

Nesse caso, o uso do dicionário pode tornar o trabalho do anotador ainda mais complicado. Em primeiro lugar, o verbo *to denounce* não evoca um cenário jurídico na FrameNet. Em segundo lugar, o verbo *to accuse* evoca um *frame* jurídico, o *frame* `Notification_of_charges`, mas esse *frame* não é apropriado para descrever o cenário jurídico evocado pela unidade lexical *denunciar* em português. Outro possível equivalente para *denunciar* poderia ser *inform against*, mas a FrameNet ainda não descreve unidades lexicais complexas de forma extensiva. Se o anotador procurar na base de dados da FrameNet por *inform*, os *frames* que ele encontrará não serão relacionados ao domínio jurídico. Esse *frames* são: `Reporting` e `Telling`.

4.2 Correspondência de *frames*

O segundo desafio que um anotador irá enfrentar ao usar as etiquetas semânticas da FrameNet é a correspondência de *frames* entre sistemas jurídicos diversos. Retomando os exemplos anteriores, as unidades lexicais *acusar* e *denunciar* têm o mesmo equivalente em inglês: *to accuse*. Na base de dados da FrameNet, a unidade lexical evoca o *frame* `Notification_of_charges`, mas os verbos *acusar* e *denunciar* são polissêmicos e podem evocar mais de um *frame* em português brasileiro. A correspondência de *frames* envolve a correspondência dos eventos jurídicos, ou seja, a correspondência de *frames* jurídicos, e a correspondência de relações entre *frames*. Assim, é necessário falar da organização do *frame* `Criminal_process` na FrameNet para se entender por que a correspondência de relações entre *frames* é tão difícil no domínio jurídico.

O *frame* `Notification_of_charges` é parte de um *frame* maior chamado `Criminal_process`. De acordo com a terminologia da FrameNet, o *frame* `Criminal_process` é um *frame* não-lexical. Esses *frames* não apresentam unidades lexicais evocadoras de *frame*. Eles representam eventos complexos, dividindo-os em *frames* mais específicos e relacionando-os por meio de relações temporais. O *frame* `Criminal_process` descreve as diferentes etapas de um processo penal de acordo com o sistema jurídico norte-americano. Na FrameNet, as relações são estabelecidas entre *frames*, não entre palavras. Assim, relações lexicais como sinonímia e antonímia não são consideradas. No caso de *frames* complexos, como o *frame* `Criminal_process`, cada sequência de eventos ou estados é descrita como um *subframe*, sendo esse *subframe* relacionado a outros *subframes* por meio da relação de precedência.

O *frame* `Criminal_process` é dividido em quatro *subframes* temporalmente relacionados por meio da relação de precedência: `Arrest`, `Arraignment`, `Trial` e `Sentencing`. O *frame* `Arraignment` é dividido em três *subframes*: `Notification_of_charges`, `Entering_a_plea` e `Bail_decision`. O *frame* `Trial` é dividido em três *subframes*: `Court_examination`, `Jury_deliberation` e `Verdict`. A figura 2 mostra a organização do *frame* `Criminal_process` e as relações entre *frames*.

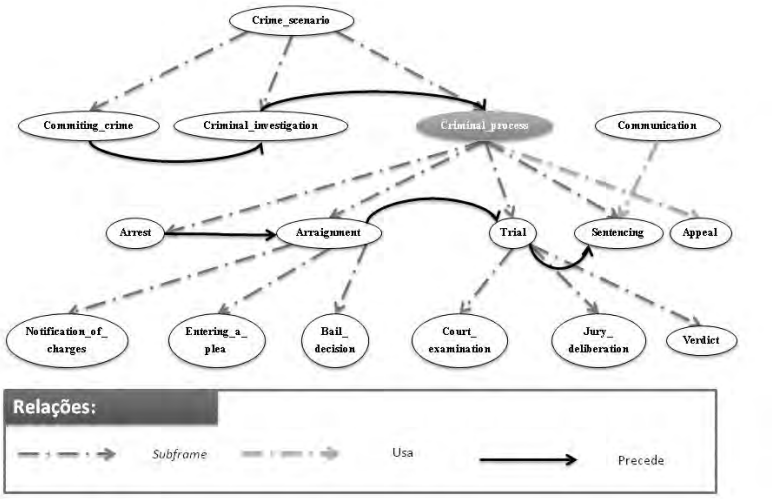


FIGURA 2 - O *frame* `Criminal_process`

O processo penal brasileiro apresenta etapas diferentes doo processo penal americano. Em primeiro lugar, um processo penal no Brasil pode seguir diferentes procedimentos. Em segundo lugar, este estudo considera apenas o procedimento especial do Tribunal do Júri. O Tribunal do Júri no Brasil é utilizado apenas em casos de crimes dolosos contra a vida. Diferentemente do processo penal americano, o processo penal brasileiro não se inicia quando o acusado é preso, mas quando a promotoria apresenta denúncia contra o acusado. Então, o juiz cita o acusado, informando-o das acusações que pesam contra ele. O acusado deve, então, apresentar a sua defesa por escrito. Essa etapa do processo penal brasileiro pode ser considerada equivalente aos *frames* `Notification_of_charges` e `Entering_a_plea`. A promotoria pode, por sua vez, apresentar uma contra-resposta à defesa do acusado e o juiz pode requerer mais investigações para esclarecer possíveis dúvidas. Após essas etapas, o acusado é levado para uma audiência preliminar. Depois da audiência preliminar, o juiz pode acolher a denúncia da promotoria. Então, o acusado é considerado réu e vai a julgamento.

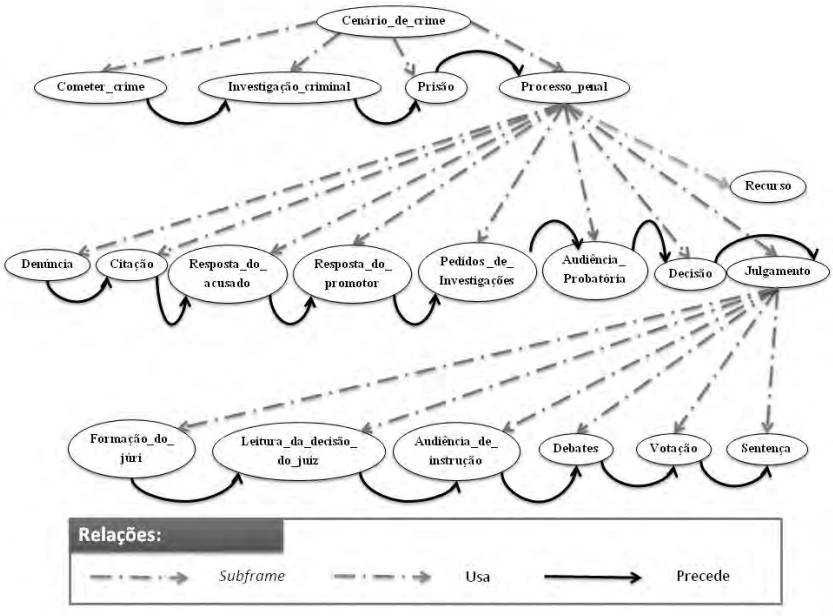


FIGURA 3 - O *frame* `Processo_penal`

Considerando os problemas de se anotar um *corpus* jurídico com etiquetas semânticas criadas para o sistema americano, decidiu-se criar *frames* que representem o sistem jurídico brasileiro para prover etiquetas semânticas para anotação de textos jurídicos brasileiros. O *frame* Processo_penal, conforme figura 3, está dividido em oito *subframes*: Denúncia, Citação, Resposta_do_acusado, Resposta_do_promotor, Pedidos_de_investigações, Audiência_probatória, Decisão e Julgamento. O *frame* Julgamento está dividido em seis *subframes*: Formação_do_júri, Leitura_da_decisão_do_juíz, Audiência_de_instrução, Debates, Votação e Sentença.

A equivalência das unidades lexicais e a correspondência dos eventos jurídicos é essencial para o uso das etiquetas semânticas da FrameNet. Apesar de a unidade lexical *acusar* ter como equivalente em inglês *to accuse*, *acusar* evoca o *frame* `Denúncia`, seu equivalente em inglês evoca o *frame* `Notification_of_charges`. Embora *to accuse* possa ser considerado como equivalente de tradução das unidades lexicais *acusar* e *denunciar*, o cenário evocado pelas unidade lexical em inglês é diferente do cenário evocado por seus equivalentes de tracução *acusar* e *denunciar*.

Considerando que um anotador procure pela unidade lexical *acusar* em um dicionário bilíngue, ele encontrará a unidade lexical *to accuse* como equivalente de tradução. Então ele procura na base de dados da FrameNet o *frame* evocado pela unidade lexical *to accuse*. Como resultado, o anotador terá os *frames* `Judgment`, `Judgment_communication` e `Notification_of_charges`. Como o anotador está anotando um *corpus* jurídico, ele pode optar pelo *frame* `Notification_of_charges` para anotar a unidade lexical *acusar*. No entanto, nem o *frame*, nem os elementos de *frame* correspondem ao cenário jurídico evocado pela unidade lexical em português. Isso gera desencontros no uso das etiquetas semânticas para anotação de textos em português.

O *frame* `Notification_of_charges` ilustra muito bem as diferenças entre os dois sistemas. Na FrameNet, o *frame* `Notification_of_charges` é a primeira etapa de um *frame* mais geral chamado `Arraignment`. A sessão de *arraignment* é uma etapa do processo penal americano que não possui correspondência no processo penal brasileiro. A segunda etapa do *frame* `Arraignment` é `Entering_a_plea` e a terceira é `Bail_decision`. Esses três *frames* estão relacionados pela relação de precedência. Portanto, se o

anotador anota a unidade lexical *acusar* como evocadora do *frame* `Notification_of_charges`, ele está anotando um texto jurídico brasileiro com informação semântica de outro sistema jurídico.

Se o anotador usa o *frame* `Criminal_process` para anotar textos jurídicos brasileiros, ele não encontrará os elementos de *frame* apropriados. Como o *frame* `Notification_of_charges` é uma etapa de um *frame* mais amplo, as relações de herança de elementos de *frame* entre os *frames* `Arraignment` e `Notification_of_charges` não são equivalentes no sistema jurídico brasileiro. Assim, em muitos casos, as relações entre *frames* da FrameNet não representam as etapas de um processo penal no Brasil. Portanto, o uso das etiquetas semânticas da FrameNet em domínios socialmente orientados, como o Direito, requer uma atenção especial para possíveis adaptações que sejam necessárias nos *frames*. Apesar de as unidades lexicais em português terem equivalentes em inglês, o cenário jurídico evocado pelas unidades lexicais em português é diferente do cenário jurídico evocado pelas unidades lexicais em inglês.

4.3 Correspondência de elementos de *frame*

Uma vez que um evento jurídico não apresenta correspondência, o *frame*, isto é, a representação esquemática de um evento jurídico, não apresentará correspondência entre dois sistemas jurídicos diferentes. Quando o cenário jurídico muda, os participantes daquele evento também podem mudar. De acordo com a definição da FrameNet para o *frame* `Notification_of_charges`, o juiz ou outra autoridade do tribunal (ARRAIGN_AUTHORITY) informa o acusado (ACCUSED) das acusaçõe (CHARGES) que pesam contra ele. Os elementos de *frame* centrais deste *frame* são ARRAIGN_AUTHORITY, ACCUSED e CHARGES.

Um anotador que esteja usando as etiquetas da FrameNet para anotar um *corpus* de textos jurídicos brasileiros terá que designar um *frame* semântico para cada unidade lexical evocadora de *frame* e elementos de *frame* para os sintagmas mais representativos de cada sentença. Ao se deparar com uma sentença cujo evocador de *frame* é a unidade lexical *acusar*, o anotador terá que identificar na FrameNet o *frame* apropriado. Como o anotador sabe que o equivalente em inglês para a unidade lexical *acusar* é *to accuse*, ele procura pela unidade lexical *to accuse* na base de dados da FrameNet e chega ao

frame `Notification_of_charges`. O anotador perceberá imediatamente que o elemento de *frame* ARRAIGN_AUTHORITY não se ajusta bem ao cenário de um processo penal no Brasil. O que o anotador deve fazer neste caso? Ele deve utilizar o elemento de *frame* ARRAIGN_AUTHORITY apesar de esse elemento não se encaixar à descrição do processo penal brasileiro ou criar um novo *frame* para descrever o cenário jurídico brasileiro evocado pela unidade lexical *acusar*?

A falta de correspondência entre os eventos jurídicos nos procedimentos penais americano e brasileiro causa a falta de correspondência entre os elementos de *frame*. Assim, o anotador terá que decidir entre usar as etiquetas da FrameNet, ignorando a falta de correspondência de *frames* e elementos de *frame*, ou criar um novo *frame* cada vez que ele perceber a falta de correspondência dos eventos jurídicos e dos papéis desempenhados pelos participantes desses eventos nos sistemas jurídicos americano e brasileiro.

4.4 Polissemia e sinonímia no domínio jurídico

A polissemia e a sinonímia são outro desafio para a anotação semântica de *corpora* jurídicos. As palavras do domínos jurídico e criminal em português podem evocar diferentes cenários, da mesma forma que unidades lexicais diferentes podem evocar o mesmo cenário jurídico. A palavra *acusar*, por exemplo, tem pelo menos duas unidades lexicais relacionadas ao domínio jurídico: *acusar* como uma unidade lexical mais geral, cujo equivalente em inglês é *to accuse*, e *acusar* como uma unidade lexical mais terminológica, cujo equivalente em inglês é *to charge*. Enquanto a unidade lexical *acusar* (*to accuse*) evoca o *frame* `Investigação_criminal`, a unidade lexical *acusar* (*to charge*) evoca o *frame* `Denúncia`. O mesmo caso de polissemia ocorre com a palavra *denunciar*. A unidade lexical *denunciar*, cujo equivalente em inglês é *to denounce*, evoca o *frame* `Investigação_criminal`. Já a unidade lexical *denunciar*, cujo equivalente em inglês é *to charge*, evoca o *frame* `Denúncia`. As palavras *acusar* e *denunciar* são sinônimas tanto em um contexto de investigação criminal (*frame* `Investigação_criminal`), quando em um cenário de denúncia de um acusado pela promotoria (*frame* `Denúncia`).

Os domínios jurídico e penal estão cheios de exemplos de palavras polissêmicas. A palavra *depor* pode evocar os *frames*

`Investigação_criminal`, `Audiência_probatória` e `Audiência_de_instrução`. A palavra *testemunhar* evoca os mesmos *frames* que a palavra *depor*. Apesar de as palavras *testemunhar* e *depor* serem sinônimas em alguns contextos, elas apresentam uma variação de significado. A palavra *testemunhar* é geralmente relacionada à testemunha de um crime que decide testemunhar voluntariamente, enquanto a palavra *depor* é usada em contextos em que a pessoa é intimada por uma autoridade a prestar testemunho. Essas diferenças de significado de uma mesma palavras podem provocar dificuldades para os anotadores.

5 Considerações finais

O estudo contrastivo demonstrou que os *frames* semânticos em domínios socialmente construídos, como o Direito, apresentam um alto grau de divergência entre as línguas. A diferença entre os sistemas jurídicos norte-americano e brasileiro faz com que os eventos jurídicos não sejam os mesmos nos Estados unidos e no Brasil. Isso provoca uma quebra entre a equivalência das unidades lexicais e a correspondência dos eventos jurídicos. Algumas unidades lexicais do inglês que apresentam equivalência em português podem não apresentar correspondência conceitual, ou seja, o significado é semelhante, mas o evento jurídico descrito por essa unidade lexical em inglês não é totalmente semelhante em português. O estudo contrastivo apontou para a necessidade de se criarem *frames* jurídicos específicos para o sistema jurídico brasileiro.

O tratamento da equivalência de *frames* é relativamente recente (LÖNNEKER-RODMAN, 2007). Em geral, os estudos de equivalência se detêm no estudo da equivalência de unidades lexicais. Os *frames* jurídicos são evidências da falta de equivalência de *frames* entre línguas, uma vez que esses *frames* representam um conhecimento socialmente construído, sendo, portanto, específicos de cada cultura e de cada país. A partir do estudo contrastivo dos *frames* semânticos da FrameNet e dos *frames* criados para o processo penal brasileiro, percebe-se que os *frames* jurídicos apresentam diferentes níveis de equivalência, variando desde a quantidade de elementos de *frame* até a natureza do evento jurídico descrito pelo *frame*. Enquanto o par de *frames* `Try_defendant` e `Julgar_acusado` possui unidades lexicais equivalentes, descrevem um evento jurídico semelhante e

apresentam os mesmos elementos de *frames* (participantes dos eventos jurídicos), o *frame* `Arraignment` não apresenta qualquer forma de equivalência.

A divergência entre os sistemas jurídicos norte-americano e brasileiro demonstrou a necessidade de criação de *frames* específicos para o processo penal brasileiro. A equivalência de *frames* entre as línguas é um fator fundamental para o sucesso no uso das etiquetas semânticas da FrameNet para anotação de *corpora* em diferentes línguas. No caso de anotação de *corpora* jurídicos, a falta de correspondência entre os sistemas jurídicos norte-americano e brasileiro traz à tona a necessidade de se criar um recurso lexical baseado em *frames* para a linguagem jurídica brasileira que possa ser utilizado para a anotação de *corpora*.

Referências

BURCHARDT, A.; ERK, K.; FRANK, A.; KOWALSKI, A.; PADÓ, S.; PINKAL, M. Using FrameNet for the semantic analysis of German: annotation, representation, and automation. In: BOAS, H. C. (Ed.). *Multilingual FrameNets in computational lexicography*: Methods and applications. Berlin/New York: Mouton de Gruyter, 2009. p. 209-244.

DINI L.; BOSCA, A. Dependency Based Valence Induction for an Italian FrameNet. In: 5TH INTERNATIONAL CONFERENCE ON GENERATIVE APPROACHES TO THE LEXICON, 2009. Proceedings... Pisa: CNR, p. 27-35.

FELLBAUM, C. A semantic network of English: the mother of all wordnets. *Computers and the Humanities*, v. 32, n. 2-3, p. 209-220, 1998.

FILLMORE, C. J. The case for case. In: BACH, E.; HARMS, R. T. (Ed.). *Universals in Linguistic Theory*. Vol. 67. New York: Holt, Rinehart and Winston, 1968. p. 1-88.

FILLMORE, C. J. An alternative to checklist theories of meaning. In: FIRST ANNUAL MEETING OF THE BERKELEY LINGUISTICS SOCIETY. *Proceedings*... Berkeley: Berkeley Linguistics Society, 1975. p.123-131.

FILLMORE, C. J. Scenes-and-frames semantics. In: ZAMPOLLI, A. (Ed.). *Linguistic Structures Processing*: Fundamental Studies in Computer Science, n. 59, North Holland Publishing, p. 55-88, 1977.

FILLMORE, C. J. Frame semantics. In: The Linguistic Society of Korea (Ed.). *Linguistics in the Morning Calm*. Seoul: Hanshin, 1982. p. 111-37.

FILLMORE, C. J. Frames and the semantics of understanding. *Quaderni di Semantica*, v. 6, n. 2, p. 222-254, 1985.

FILLMORE, C. J. Valency Issues in FrameNet. In: HERBST, T.; GÖTZ-VOTTELER, K. (Ed.). *Valency*: Theoretical, Descriptive and Cognitive Issues. Berlin, New York: Mouton de Gruyter, 2007. p. 129-160.

FILLMORE, C. J.; BAKER, C. A frames approach to semantic analysis. In: HEINE, B.; NARROG, H. (Ed.). *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, 2010. p. 313-339.

FILLMORE, C. J.; JOHNSON, C. R.; PETRUCK, M. R. L. Background to FrameNet. *International Journal of Lexicography*, v. 16, n. 3, p. 235-250, 2003.

GOFFMAN, E. *Frame Analysis*. New York: Harper, 1974.

GOYOS Jr., D. N. *Noronha's Legal Dictionary – Noronha Dicionário Jurídico: English-Portuguese, Portuguese-English – Inglês-Português, Português-Inglês*. 1. ed. São Paulo: Observador Legal, 1992.

LÖNNEKER-RODMAN, B. *Multilinguality and FramNet*. Technical Report. TR-07-001. Berkeley: ICSI, 2007.

MILLER, G. A. WordNet: a lexical database for English. *Communications of the ACM*. New York: ACM Press, v. 38, n. 11, p. 39-41, 1995.

MINSKY, M. *A framework for representing knowledge*. Artificial Intelligence Memo Nº 306. Cambridge, MA: Massachusetts Institute of Technology, 1974.

PADÓ, S. *Cross-lingual Annotation Projection Models for Role-Semantic Information*. PhD Thesis. Saarbrücken: Universität des Saarlandes, 2007.

PADÓ, S.; LAPATA, M. Cross-lingual projection of role-semantic information. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE AND CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE. *Proceedings*... Vancouver: Association for Computational Linguistics, p. 859-866, 2005.

PADÓ, S.; PITEL, G. Annotation précise Du français en sémantique de rôles par projection cross-linguistique. In: TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES. *Proceedings*... Toulouse, France, 2007.

TESNIÈRE, L. *Éléments de syntaxe structurale*. Paris: Klincksieck, 1959.

TONELLI, S.; PIANTA, E. Frame Information Transfer from English to Italian. In: THE SIXTH INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION (LREC'08). *Proceedings*... Marrakech: European Language Resources Association (ELRA), p. 28-30, 2008.

VENTURI, G.; LENCI, A.; MONTEMAGNI, S.; VECCHI, E.; SAGRI, M.; TISCORNIA, D.; AGNOLONI, T. Towards a FrameNet Resource for the Legal Domain. In: THIRD WORKSHOP ON LEGAL ONTOLOGIES AND ARTIFICIAL INTELLIGENCE TECHNIQUES, WORKSHOP ON SEMANTIC PROCESSING OF LEGAL TEXTS (LOAIT '09). *Proceedings*... Barcelona, Spain. 8 June, 2009.

Superando o estado da arte na etiquetagem morfossintática por meio de regras de pós-etiquetagem

Cid Ivan da Costa Carvalho[1]
Davis Macedo Vasconcelos[2]
Leonel Figueiredo de Alencar[3]

RESUMO: Segundo Kveton e Oliva (2002), os erros de etiquetagem, quando se utiliza um *corpus* para treinar estatisticamente algum sistema de processamento automático da linguagem natural, constituem desvios das regularidades que se espera que o sistema aprenda, resultando num modelo falso da língua. Esses autores propõem um método de correção desses erros num processo de pós-etiquetagem, levando em conta a detecção de *n*-gramas impossíveis na língua a ser modelada. Neste trabalho, procura-se sistematizar os erros cometidos pelo etiquetador morfossintático Aelius, construído por meio do NLTK (BIRD, KLEIN e LOPER, 2009). Como ponto de partida deste trabalho, compilamos um *corpus* de textos de comunicação mediada por computador, gênero que se caracteriza pela abundância de abreviaturas, grafias não padrão e desvios da norma culta. Após a anotação feita com o Aelius, levantamos os erros cometidos pela ferramenta e os classificamos. Com base nisso, elaboramos um primeiro conjunto de regras para correção automática da etiquetagem. Numa segunda etapa, tendo como meta chegar a 99% de acurácia, acima, portanto, do estado da arte de 97%

[1] Professor Assistente da Universidade Federal Rural do Semi-Árido e doutorando pela Programa de Pós-Graduação em Linguística pela UFC. cidivanc@gmail.com.

[2] Professor do Instituto Federal do Ceará e doutorando pela Programa de Pós-Graduação em Linguística pela UFC,. davis.ifce@gmail.com.

[3] Professor e orientador do Programa de Pós-Graduação em Linguística e do Departamento de Letras Estrangeiras da Universidade Federal do Ceará. leonel_de_alencar@yahoo.com.br.

(GÜNGÖR, 2010, p. 207), implementaremos regras em Python para correção dos erros cometidos por esse sistema de etiquetagem.
PALAVRAS-CHAVE: Aelius, erros de etiquetagem, *n*-grama, *corpus*.

ABSTRACT: According to Kveton e Oliva (2002) tagging errors, when a corpus is used in order to statistically train a natural language processing system, constitute deviations from regularities which the system is expected to learn, leading to a false model of the analyzed language. These authors propose a correction method for these errors in a post-tagging process, taking into account the identification of impossible n-grams in the language to be modeled. In this paper we try to systematize the errors made by the Brazilian Portuguese morphosyntactic tagger Aelius, built through the NLTK (BIRD, KLEIN and LOPER, 2009). The departing point was the compilation of computer mediated communication corpus. This genre is characterized by an abundance of abbreviations, spelling variation and standard language deviations. After the Aelius annotation was done, we listed the errors made and classified them. This was the basis for a set of rules for automatic tagging correction. Our goal was to reach a 99% accuracy, above the state of the art 97 (GÜNGÖR, 2010, p. 207), and in order to reach that we implement a system of Python rules to correct the errors created by the tagging system.
KEYWORDS: Aelius, tagging errors, n-grams, *corpus*.

1 Introdução

Antes de desenvolver os programas computacionais no universo linguístico é preciso "procurar as forças que estão em jogo, de modo permanente e universal. Em todas as línguas e deduzir as leis gerais as quais se possam referir todos os fenômenos peculiares" (SAUSSURE, 2006, p. 24) – ou seja, identificar "uma estrutura linguística imutável que sustenta a língua e subjaz a quaisquer outras realizações que dela se façam" (SAUTCHUK, 2010, p.3).

Esse ponto é fundamental para reiteração do caráter científico da linguística. Mioto (2007, p.13) enuncia que "Também na linguística esperamos ser capazes de fazer observações atentas e acuradas de maneira tão objetiva e imparcial quanto possível". É uma busca pelos "princípios que estejam na base de todo fenômeno sintático existente.

A esse conjunto de postulações básicas e de afirmações consequentes chamamos um modelo teórico.

É através dessa colaboração entre ciência da computação, através da subdisciplina da Inteligência Artificial - IA, e a linguística consolidar-se-ia mais tarde numa nova disciplina, a linguística computacional." (ALENCAR, 2006, p. 12-13) – ciência, hoje, subdividida em "linguística de corpus e o processamento de linguagem natural (PLN)". (OTHERO; MENUZZI, 2005, p. 22).

Nesse contexto, a linguística computacional caracteriza-se pela utilização de "computadores para o armazenamento e acesso a textos escritos ou falados" que "pode ser rapidamente pesquisado para informações a respeito da regularidade da língua, tais como frequência de palavras, de formas ou de construções". (VIEIRA; LIMA, on-line). Sua relevância é verificada na "elaboração de teorias gramaticais formalmente mais consistentes e psicolinguisticamente mais realistas […] e, assim, testar, com um grau de sofisticação que dificilmente poderia ser atingido por seres humanos, a adequação dos modelos postulados". (ALENCAR; OTHERO, 2011, p. 9-10).

Uma moderna definição para corpus linguístico é:

> uma coleção classificada de objectos linguísticos para o uso em Processamento de Linguagem Natural / Linguística Computacional /Linguística em que uso pode ser estudo, medição, teste, ou avaliação enquanto variados *objectos linguísticos* são textos, frases, palavras, entrevistas, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, filmes, imagens com legendas, traduções, correções (de textos de alunos de língua ou de tradução), telefonemas, simulações do tipo Wizard of Oz, programas (SANTOS, 2008, p. 45-46).

Esses dois conceitos acima expressos estão relacionados com a ferramenta de uso para o processamento da linguagem natural. A ferramenta, *Natural Language ToolKit* (doravante NLTK), aqui utilizada. Ela facilita a manipulação ou tratamento textual no que se refere ao processo de etiquetagem sintática – ou seja, categorização de palavras em classes gramaticais: substantivos, verbos, adjetivos, advérbios (no idioma Inglês). Tal ferramenta é utilizada pelo etiquetador Aelius.

Esse artigo se divide em quatro partes: no primeiro, apresenta-se a arquitetura do sistema Aelius, as funções para o pré-processamento

de corpora, o processo de anotação sintática e o tagset do sistema; no segundo tópico, expõe alguns erros de etiquetagem mais frequentes e relevantes cometidos por esse pacote em Python, feita com textos mediados por computador, bem como, um exemplo de implementação feita em Python com o intuito de aumentar da acurácia.

2 Etiquetador Aelius

A etiquetagem morfossintática das partes do discurso é uma das tarefas mais estudadas no processamento de linguagem natural. Ela se constitui como uma área muito importante, uma vez que auxilia em algumas ferramentas e em modelos de implementação que leva a aplicação em determinadas tarefas. Além disso, é uma tarefa aparentemente simples para o processamento da linguagem, no entanto, o desempenho de outras ferramentas depende diretamente desse processo, como mencionados acima.

Além disso, a etiquetagem morfossintática é uma tarefa intermediária que tem como objetivo principal analisar e entender a língua natural. Porém, a maior dificuldade nesse sistema reside na ambiguidade das palavras, em que cada palavra pode pertencer a várias categorias discerníveis geradas pelo contexto linguístico e/ou nos módulos desenvolvidos pelo sistema.

O processo de anotação automática é feito pelos etiquetadores morfossintáticos. Para a etiquetagem dos textos mediados por computador, utilizou-se o sistema Aelius.

Para compor o *corpus* de texto, selecionou-se comentários de blogs com temáticas sobre a educação. Esse gênero possui uma característica muito peculiar: a espontaneidade no uso linguagem, sendo assim, mais informal, o que contribui para observação do desempenho de na etiquetagem morfossintática.

O etiquetador Aelius[4] é uma ferramenta para anotação automática de *corpora* que possui uma arquitetura híbrida, ou seja, "recorre às regras, formuladas manualmente em expressões regulares, para etiquetar as palavras inexistentes no *language model*", (ALENCAR, 2010, p. 3), e ao sistema estatístico estocástico baseado

[4] O sistema Aelius é livremente disponível no endereço http://sourceforge.net/projects/aelius/files/.

em *n*-grama. Ele desempenha o pré-processamento de *corpus* de treino e anota o texto, com o *tagset* do Tycho Brahe, verificando o uso de nomes próprios no contexto linguístico .

As regras formuladas contribuem para a resolução de ambiguidade na etiquetagem. Segundo Voutilainen (2009), essas regras podem ser baseados sobre duas fontes de informações, ambas codificada no etiquetador na forma de uma linguagem modelo: a informação sobre a palavra em si, ou seja, em que contexto efetivo a palavra é mais usado, por exemplo, a palavra como: verbo ou advérbio; e as informações sobre a sequência palavra/etiqueta (ou contexto informacional): isto é o modelo pode preferir analisá-lo como um verbo a uma conjunção, se o termo precedente for uma advérbio ou um determinante.

O componente estatístico contribui durante o treinamento do sistema, produzindo regras simbólicas que, quando utilizada no corpus de teste, podem ser modificadas. Nas palavras de Voutilainen (2009), o sistema estatístico é, muitas vezes, baseada no que é conhecido sobre o léxico. Se, por exemplo, é sabido que o léxico contém todas "as classes fechadas de palavras" assim como os pronomes e artigos, o "adivinhador" pode com segurança propor apenas a análise de classes abertas (isto é, a análise de nomes e verbos).

Além de sua arquitetura, o Aelius possui uma série de funções em Python que fazem o pré-processamento do *corpus* de treino, conforme parâmetros especificados pelo usuário; nele, o usuário pode, também, especificar um conjunto de etiquetas a serem ignorados na construção do modelo, bem como, dividir o *corpus* de base em um número específico de blocos e embaralhá-los de forma aleatória, o que permite obter um *corpus* de treino e um corpus de teste mais balanceados. (ALENCAR, 2010).

Outro módulo fundamental desse sistema é a anotação morfossintática. Antes da anotação, no *input*, o Aelius realiza o processo de tokenização, ou seja, o interior do texto é dividido em *tokens* para outra análise como: marcas de pontuação, palavras compostas unidas por hífens, etc.

Depois a anotação de *corpus* nessa ferramenta ocorre da seguinte forma: no *input*, entra-se um texto não etiquetado. Enquanto etiqueta o texto, o sistema verifica se a palavra pertence a classe dos nomes próprios, nos casos afirmativos, o sistema etiqueta a palavra com a *tag* NPR. Depois, no *output*, o Aelius retorna um texto

etiquetado com o *tagset* do Corpus Histórico do Português Tycho Brahe.[5] Como mostra a figura abaixo:

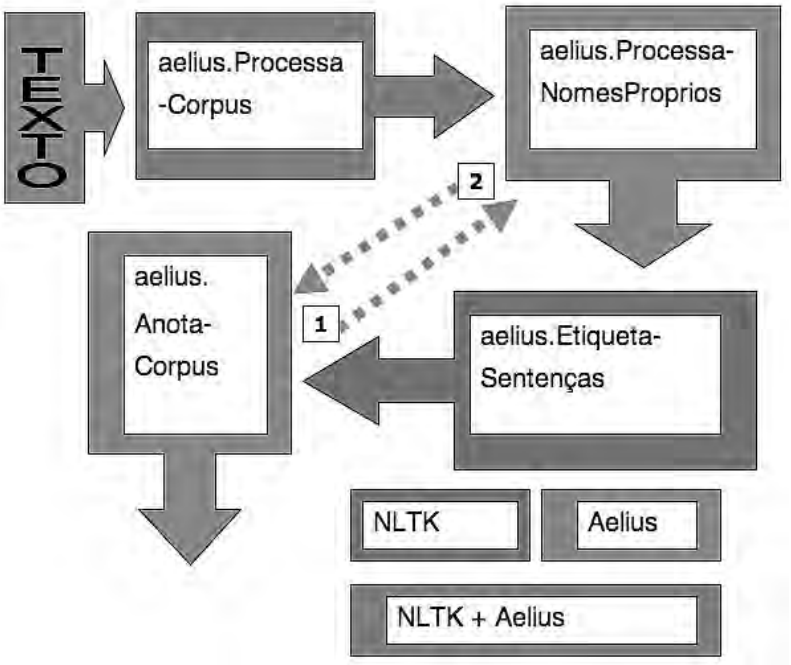


FIGURA 1
Anotação morfossintática pelo Aelius
Fonte: Alencar (2010)

Alencar (2010, p. 4), referindo ao módulo de nome próprio, diz que:

> A grande quantidade de erros relacionadas à etiqueta NPR (nome próprio) levou-nos a desenvolver e implementar em Python um algoritmo para distinguir nomes próprios de outras palavras com inicial maiúscula. Esse algoritmo supera grave deficiência do etiquetador de expressões regulares do NLTK, que não leva em conta o contexto onde um *token* ocorre.

[5] Para maiores informações sobre o *Corpus* Histórico do Português Tycho Brahe consulte o sítio do projeto: http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html.

Esse módulo desenvolvido em Python contribui para o aumento da acurácia desse sistema, mesmo utilizando o *tagset* extenso como o do Tycho Brahe, o Aelius é ferramenta dispõe de alta acurácia na anotação morfossintática para textos literários do século XIX, com desempenho de 95% de acertos.

Corpus Histórico do Português Tycho Brahe foi desenvolvido junto ao projeto Padrões Rítmicos, fixação de parâmetros e mudanças linguísticas; é um *corpus* eletrônico anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1845.

Atualmente, possui 53 textos (2.464.191 palavras) que estão disponíveis para pesquisa livre, com um sistema de anotação linguística em duas etapas: anotação morfológica (aplicada em 31 textos); e anotação sintática (aplicada em 14 textos).

Da organização desse projeto, foi elaborado um conjunto de etiquetas adequadas à representação e à discriminação das categorias necessárias à descrição dos enunciados da língua portuguesa para aquele período. Para tanto, dividiram o sistema de etiquetas morfológicas em dois grupos que, ao todo, perfazem um total de 383 *tags*, segundo Milidíu, Santos e Duarte (2008): as etiquetas categorias e as etiquetas flexionais. As primeiras classificam o item lexical em uma classe gramatical e as segundas indicam traços como gênero, número, pessoa, tempo e modo (CORPUS, 2010). Exemplo, desses grupos, podemos ver nos verbos plenos que recebem a etiqueta VB, quando está no infinitivo e; quando flexionais, acrescenta- se a essa, a etiqueta que determina a flexão do verbo. Dessa maneira, o tagset do corpus Tycho Brahe procura representar os termos das sentenças, tendo em conta a classificação das palavras e as discrimina conforme a flexão verbal e/ou nominal das unidades do discurso.

Como visto acima, o Aelius é um sistema com grande desempenho para a etiquetagem do português brasileiro, especialmente, para textos literários do século XIX. Porém, com o intuito de melhorá-lo no que se refere à precisão para os textos atuais, apresentam-se alguns erros mais frequentes cometidos por esse sistema quando testado em corpus de textos mediados por computador.

3 Erros e correção de pós-etiquetagem

A correção manual de textos automaticamente etiquetados levou a perceber certos erros frequentes na etiquetagem. Estes erros,

muitas vezes, eram claramente dependentes do contexto linguístico em que as palavras ocorriam. Em português, os nomes próprios e nomes comuns podem vir depois de artigos definidos, porém o etiquetador cometeu claramente um erro nos exemplos a seguir:

(1) … o/D professor/NPR que/C não/NEG adotar/VB o/D livro/N didático/ADJ...

(2) Espera/VB-I -/+ se/SE que/C o/D professor/NPR tire/VB-SP...

pois o termo "professor" foi etiquetado como nome próprio no (1) e no (2), enquanto deveria ser etiquetado como nome, ou seja, receber a etiqueta N. Esse termo está precedido de um determinante. Diferentemente do que acontece com os exemplos acima, o contexto (3) apresenta a etiquetagem, com a mesma palavra, se deu de forma correta.

(3) … muitos/Q-P professores/N-P fazem/VB-P um/D-UM

Outro caso de erro em que se dever observar com muita atenção foi a etiquetagem feita com os termos "Escola e escolar" que receberam, inadequadamente, as etiquetas NPR e VB, respectivamente, porém deveriam ser etiquetados como sendo nome/N e adjetivo/ADJ-G. Veja o contexto desse termos nas expressões em que foram utilizadas no *corpus*:

(4) … a/D-F realidade/N escolar/VB que/C dirigem/VB-P .

(5) … aconteceu/VB-D na/P+D-F Escola/NPR que/WPRO trabalhava/VB-D

No contexto (4), o sistema reconheceu a palavra "escolar" como sendo um verbo no infinitivo, logo, recebendo a etiqueta VB. Esse erro foi cometido devido à palavra terminar com AR, que é uma desinência dos verbos no infinitivo e, além disso, tanto a classe dos verbos quanto a classe dos adjetivos podem vir depois de um Nome. Esse problema não pode ser corrigido com um algorítmico que leva em consideração o final do termo AR antecedido de um nome/N deveria ser etiquetado com ADJ-G, pois os verbos também podem ser precedidos de nome.

O erro no contexto (5) não constitui inconsistência no processador de nomes próprios implementados (ver ALENCAR, 2010, p.5) ocorrido nessa situação, pois outros termos que inicia com letra maiúsculas não são etiquetados com NPR. Outros exemplos, no mesmo *corpus*, apresentam que o módulo implementado é bastante consistente. Veja os exemplos:

(6) … este/D assunto/N nas/P+D-F-P escola/N municipais/ ADJ-F-P …

(7) Nós/PRO em/P São/NPR Caetano/NPR do/P+D Sul/ NPR queremos/VB-P …

No contexto (6), a mesma palavra foi etiquetada corretamente. Observe ainda que o termo está antecedido pela etiqueta P+D, ou seja, pronome e determinante. Da mesma forma no (7), o termo "São" poderia ser etiquetado como sendo verbo ser, porém foi classificado como sendo um nome próprio. O erro no exemplo (5) foi cometido devido a erros no corpus de teste.

Além desse casos, percebem-se outras inconsistências na etiquetagem de termos em início de sentenças. Veja os exemplos (8) e (9).

(8) Envio/VB-P de/P verbas/N-P para/P este/D fim/N …

(9) Concordo/N totalmente/ADV contigo/P+PRO ./.

A palavra "Envio", que é um nome, foi etiquetada como sendo VB-P. De modo inverso acontece com o termo "Concordo" que é VB-P e está etiquetado como sendo nome. Esse caso é necessário observar que as duas palavras apareceram no início da sentença com as letras em maiúsculas. A etiquetagem da primeira palavra não levou em consideração a preposição que com ela forma um bigrama.

Para título de exemplo, foi feito um módulo Python de para correção desses erros dentro do contexto apresentado.

4 Implementação em Python

Abaixo, apresentam-se os processos pelos quais foram realizados o estudo: o programa 1 realiza etiquetagem com o sistema

Aelius, o programa 2, dada uma sentença, faz a correção automática dos erros em bigrama e o programa 3 verifica a correção das sentenças.

Em Python o símbolo sustenido (#) indica que o resto da linha é um comentário. O extensão do arquivo de ser em texto puro, ou seja, em txt para que o sistema possa processar os dados.

Programa 1: Etiquetagem pelo Aelius

```
# define o arquivo de entrada
arq = "ENTRADA.txt"

# importa o projeto Aelius
import Aelius

# realiza etiquetagem, gerando o arquivo anotado
from Aelius import AnotaCorpus
AnotaCorpus.anota_texto(arq,"AeliusBRUBT.pkl")

# renomeia arquivo temporário
import os
os.rename("ENTRADA.nltk.txt","SAIDA.etiquetado.txt")
```

Programa 2: Correção automática de sentenças

```
# processa as sentenças
tam = len(linhas)
# processa as sentenças (sents)
for i in range(0,tam):
sent = linhas[i]
sents = sent.decode("utf-8").split()
for cad in sents:
y = cad.find("@") # verifica se houve correção, ou seja, se existe o simbolo "@"
if (y > -1):
```

```
x = cad.find("/") # se houve, procura o simbolo "/"
palavra = cad[0:x] # copia para "palavra" o substring antes do "/"
nova_etiq = cad[y+1:] # copia para "nova_etiq" o substring depois
do "@"
out = palavra + "/" + nova_etiq # ajusta "nova_etiq" para "out"
else:
out = cad # se não houve correção
lista.append(out) # "out" é inserida na lista
```

Programa 3: Verificação da correção

```
sent='que/C o/D professor/NPR tire/VB-SP'
sents=sent.decode('utf-8').split()
from nltk import bigrams
cad=bigrams(sents)
for(w1,w2)in cad:
        x=w1.find('/')
        palavra1=w1[0:x]
        tag1=w1[x+1:]
        x=w2.find('/')
        palavra2=w2[0:x]
        tag2=w2[x+1]
        if ((tag1=='D')and(palavra2=='professor')):
                tag2='N'
                print palavra2,'/',tag2
```

A correção das outras sentenças segue o mesmo exemplo da primeira, pois a implementação segue uma relação de bigramas e não uma relação entre etiquetas, ou seja, o contexto em que os termos estão dispostos na sentença. Isso produz um ganho na acurácia do etiquetador, uma vez que trata o problema local, mas contextos com outras palavras.

5 Considerações

Alencar (2010), no desenvolvimento do etiquetador, procedeu incrementalmente, repetindo várias vezes o ciclo editar – compilar – testar – depurar o sistema de etiquetagem. Esse processo envolveu contribui para otimização do etiquetador baseado em expressões regulares e, também, a eliminação de tokens com etiquetas que não se referem à análise morfossintática bem como correções de inconsistências do próprio *corpus* de treino.

Ele acrescenta que a grande quantidade de erros relacionadas à etiqueta NPR (nome próprio) levou a desenvolver e implementar em Python um algoritmo para distinguir nomes próprios de outras palavras com inicial maiúscula. Esse algoritmo supera grave deficiência do etiquetador de expressões regulares do NLTK, que não leva em conta o contexto onde um *token* ocorre. Todavia, é necessário depurar e rever o módulo para aumentar a qualidade na etiquetagem, quando o *corpus* é constituído por textos atuais.

Além disso, outra possibilidade de aumentar a acurácia do etiquetador é diminuir erros que envolvem contextos sintáticos claramente identificáveis, por meio da inclusão de regras de reetiquetagem elaboradas manualmente.

Referências

ALENCAR, Leonel Figueiredo de. Aelius: uma ferramenta para anotação automática de corpora usando o NLTK. IX ENCONTRO DE LINGUÍSTICA DE CORPUS. *Anais*... Porto Alegre, PUCRS, 8 e 9 de outubro de 2010.

ALENCAR, Leonel Figueiredo de. *Linguagem e Inteligência Artificial*: na coletânea Linguagens. As expressões do Múltiplo. Fortaleza: Premius, 2006.

ALENCAR, L. F.; OTHERO, G. A. (Org.). *Abordagens computacionais da teoria da gramática.* Campinas, SP: Mercado das Letras, 2011.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*: Analyzing Text with the Natural Language Toolkit. Creative Commons Attribution-NonCommercial-No Derivative Works 3.0. New York, 2007.

CARNIE, A. *Syntax*: a generative introduction. Blackwell Publishing, 2007.

CORPUS Histórico do Português Tycho Brahe. Campinas: Instituto de Estudos da Linguagem/Universidade Estadual de Campinas, 2010. Disponível em: <http://www. tycho.iel.unicamp.br/~tycho/corpus/ > Acesso em: 30. set. 2010.

KVETON, P.; OLIVA, K. (Semi-)Automatic Detection of Errors in PoS-Tagged Corpora. INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. *Proceedings*... n. 19, 2002, Taipei. Stroudsburg: Association for Computational Linguistics, 2002.

MILIDIU, Ruy Luiz; SANTOS, Cícero Nogueira dos; DUARTE, Julio Cesar. *Portuguese corpus-based learning using ETL*. J. Braz. Comp. Soc. [online], v. 14, n. 4, p. 17-27. ISSN 0104-6500, 2008.

MIOTO, C.; SILVA, M. C. F.; VASCONCELLOS, R. E. *Novo manual de sintaxe*. Florianópolis: Insular, 2007.

OTHERO, G. A; MENUZZI, S. M. *Linguística computacional*: teoria & prática. São Paulo: Parábola Editorial, 2005.

SANTOS, D. Corporizando algumas questões. In: TAGNIN, Stella E. O.; VALE , Oto Araújo. (Org.). Na coletânea *Avanços da linguística de corpus no Brasil*. São Paulo: Humanitas, 2008.

SARDINHA, T. B.; ALMEIDA, G. M. B. A Linguística de Corpus no Brasil. In: TAGNIN, Stella E. O.; VALE , Oto Araújo. (Org.). Na coletânea *Avanços da linguística de corpus no Brasil*. São Paulo: Humanitas, 2008.

SAUTCHUK, I. *Prática de morfossintaxe*: como e por que aprender aprender análise (morfo)sintática. 2. ed. Barueri, SP: Manole, 2010.

SAUSSURE, F. *Curso de Linguística Geral.* 27. ed. São Paulo: Cultrix, 2006.

VIEIRA, R.; LIMA, V. L. *Linguística computacional*: princípios e aplicações. São Leopoldo: Unisinos. Disponível em: www.inf.unioeste.br/~jorge/.../linguística%20computacional.pdf.

VOUTILAINEN, Atro. Part-of-speech tagging. In: MITKOV, Rusland (Org.). *The Oxford Handbook of computational linguistics.* Oxford: Oxford university Press, 2009.

Um paralelo entre o *frame* de comunicação do português e do inglês

Francine Ferreira Vaz[1]
Luiz Fernando Matos Rocha[2]

RESUMO: Essa pesquisa integra o projeto *FrameNet* Brasil que tem como um dos seus objetivos definir, através da análise de *corpus*, os *frames* da Língua Portuguesa baseados naqueles da Língua Inglesa descritos pelo projeto *FrameNet* americano. Nessa perspectiva, focamos o nosso trabalho nos *frames* de Comunicação, Meios de Comunicação e Modos de Comunicação do Português do Brasil. Esse estudo tem como suporte teórico a Semântica de *frames* (FILLMORE, 1982) e foi desenvolvido nos moldes do Projeto *FrameNet* americano, partindo da descrição dos *frames* e da definição dos seus elementos centrais e periféricos para, a seguir, anotar as Unidades Lexicais (UL) de cada um nos *corpora* escolhidos. Dessa forma, foi possível criar tabelas com os padrões encontrados e compará-los com os padrões originais do inglês. A hipótese inicial de que os *frames* descritos para o português seguiriam os moldes dos *frames* descritos para o inglês foi comprovada; no entanto, percebeu-se que o comportamento das ULs de cada *frame* é um pouco distinto nas duas línguas estudadas. Analisando esses três *frames*, esse trabalho contribui para a descrição dos *frames* para o Português e para que, futuramente, seja possível criar uma rede de equivalentes semânticos entre todas as línguas que forem descritas dessa forma.

PALAVRAS-CHAVE: *Frame*, *FrameNet*, *Corpus, Frame* de Comunicação,*Frame* de Meios de Comunicação, *Frame* de Modos de Comunicação

[1] Doutoranda em Linguística pela UFJF. Email: franfv@uol.com.br.

[2] Professor do Programa de Pós-Graduação em Linguística da UFJF. Email: luiz.rocha@ufjf.edu.br.

ABSTRACT: This research is part of a larger project called *FrameNet* Brasil that, among other aims, intends to, through corpus analyses, define the Portuguese frames based on the English ones described by the Berkeley FrameNet Project. Thereby, this work focuses on the description of the following Portuguese frames: Communication, Communication Means and Communication Manner. This study is supported by the frame semantics theory and was developed according to the Berkeley FrameNet Project model. First, the frames were described and the core and non-core elements were defined. Then, the lexical units were annotated in the corpora chosen. By doing this, it was possible to create tables with the patterns found and compare then to the English original. The initial hypothesis that the Portuguese frames are similar to the English ones was confirmed, even though, we noticed that the ULs behave differently in these two languages. Analysing these three frames, this work contributes to the description of the Portuguese frames and to the creation, in the future, of a network with semantic equivalents among languages based on frame.

KEYWORDS: Frame, FrameNet, Corpus, Communication Frame, Communication_means frame and Communication_manner frame

1 A semântica de *frames*

O termo "Semântica de *Frames*" se refere a um programa de pesquisa em semântica empírica e a um *framework* descritivo que será usado para representar o resultado dessa pesquisa. A Semântica de *Frames* oferece uma nova forma de postular princípios para a criação de novas palavras e sintagmas, para acréscimo de sentido às palavras já existentes e para a junção dos significados dos elementos de um texto de acordo com o sentido geral deste. Já o termo *frame* se refere a um sistema de conceitos relacionados de tal maneira que, para compreender qualquer um deles, é preciso entender toda a estrutura na qual ele se encaixa; quando um dos elementos dessa estrutura é introduzido no texto ou na conversação, todos os outros ficam automaticamente disponíveis. Esse termo é usado para englobar um conjunto de conceitos presentes na literatura de compreensão de linguagem natural, como esquema, *script*, cenário, modelo cognitivo e outros (FILLMORE, 1982).

No desenvolvimento das descrições da semântica de *frames*, é necessário identificar, primeiramente, fenômenos, experiências ou

cenários, representados pelo significado das palavras alvo, e as sentenças nas quais elas ocorrem. Depois, identificam-se e criam-se etiquetas para aquelas partes ou aspectos associados a sentidos específicos das expressões linguísticas. Assim, definimos os elementos de *frame*.

Um elemento de *frame* é simplesmente um participante regular, característica ou atributo do tipo de situação descrita pelo *frame*. Assim elementos de *frame* de casamento incluem, por exemplo, noivo, noiva, padrinho, madrinha, cerimônia. Elementos de *frame* não são obrigatórios, existe casamento sem padrinho, mas precisam ter características regulares recorrentes.

Em semântica de *frames*, todos os sentidos das palavras são relativos ao *frame,* no entanto, o sentido de uma palavra não ativa um *frame* inteiro. Palavras diferentes selecionam diferentes aspectos do contexto de um perfil. Algumas vezes, esses aspectos são apenas mutuamente exclusivos devido ao tipo de circunstância descrita para participantes distintos, como mulher e homem em um *frame* de casamento. No entanto, alguns significados das palavras não se diferenciam no que eles perfilam, mas sim em como eles perfilam. Em alguns casos, as palavras se diferenciam na perspectiva.

Consideremos o verbo "arriscar", discutido em Fillmore e Atkins (1998), que permite vários tipos de participantes dentro de um único espaço gramatical.

a) John arriscou ser repreendido.

b) John arriscou seu carro.

c) John arriscou um passeio na pista de esqui avançada.

O *frame* do verbo "arriscou" tem três participantes diferentes, (a) uma coisa ruim que pode acontecer, (b) uma coisa valiosa que pode ser perdida e (c) uma atividade que pode levar ao acontecimento de uma coisa ruim. Todas podem acontecer na posição de objeto direto, como mostrado acima. Como existem três relações diferentes, uma teoria que associasse significados lexicais a relações marcaria três sentidos diferentes. No entanto, a semântica de *frames* descreve como um único *frame* com três perfilamentos diferentes, o que se torna possível graças à associação da estrutura do *frame* com as opções de perfilamento que a língua oferece.

Fillmore (1992) afirma que os dicionários padrões não estão equipados para apresentar uma organização polissêmica, porque eles não oferecem um meio de acessar detalhes de *frames* conceptuais dados. Uma representação mais apropriada pode ser oferecida em um dicionário digital baseado em *frame* com as propriedades citadas acima.

2 A *FrameNet*

O projeto *FrameNet*, em desenvolvimento desde 1997, é liderado pelo Professor Charles Fillmore, no *International Computer Science Institute* (ICSI), em Berkeley, na Califórnia. De acordo com Ruppenhofer *et al*. (2010), o objetivo do projeto é criar "um recurso lexical *on-line* baseado na semântica de *frames* e suportado por evidência de *corpus*". Dessa forma, torna-se possível documentar as possibilidades semânticas e sintáticas de cada palavra (valências) e de cada sentido dessa palavra através da anotação de frases exemplares e análise de resultados.

Segundo os dados disponíveis no site oficial do projeto (www.framenet.icsi.berkeley.edu), o banco de dados já contém mais de dez mil unidades lexicais e mais de 960 *frames* anotados e exemplificados através de 170.000 sentenças. Esses dados, referentes somente à língua inglesa, são liberados ao público (já está na terceira versão) e utilizados também por outros pesquisadores que estão ampliando esse projeto para outras línguas como o espanhol, o alemão, o chinês, o japonês e o português (Projeto *FrameNet* Brasil liderado pela professora Margarida Salomão, na UFJF, desde 2009). Segundo Salomão (2009, p.5),

> Na conclusão deste trabalho, estaríamos nos aproximando do sonho do "dicionário ideal", no qual cada um de nós, ao consultar uma palavra, seríamos remetidos imediatamente para o *frame* que ela evoca, com todos os respectivos Elementos componentes; veríamos, além disso, uma listagem de todas as valências desta palavra, suas possibilidades combinatórias sintáticas e semânticas, ilustradas por exemplos correspondentes. A consulta ainda nos ofereceria um conjunto de outras palavras que evocam o mesmo *frame* e o conectaria com outros *frames* semanticamente relacionados.

Uma outra aplicação desse projeto é auxiliar o trabalho de rotulação semântica de texto corrido para aplicação no Processamento

de Linguagem Natural, o que facilitaria a comunicação homem-máquina, e geraria uma grande transformação no modo como a informação é armazenada e acessada no mundo tecnológico. O exemplo mais claro disso seria uma maior eficiência nas buscas eletrônicas, já que a *web* deixaria de ser organizada sintaticamente (grande quantidade de informação e busca com resultados insatisfatórios) e passaria a ser organizada semanticamente, usando mecanismos capazes de capturar o significado das informações.

Existem duas categorias que são frequentemente usadas no Projeto e por isso necessitam de ser definidas mais objetivamente: as unidades lexicais e as anotações. Entende-se por unidade lexical (LU, *lexical unit* em inglês) o par formado por uma palavra e o seu significado. Cada significado diferente de uma palavra faz com que ela seja associada a um *frame* diferente. O conceito de *frame* define uma estrutura conceitual que descreve uma situação em particular, objetos e eventos com seus participantes e propriedades. Os *frames* podem ser evocados tanto por substantivos, adjetivos, quanto por advérbios e preposições. Por exemplo, o *frame* de comunicação descreve uma ação que envolve um Comunicador que transmite uma Mensagem a um Destinatário e é evocado por palavras como destinatário, tema, assunto, tópico. Os elementos que participam desse *frame* (comunicador, destinatário, mensagem e outros) são chamados de elemento de *frame* (EF).

Em relação às anotações, de acordo com Ruppenhofer *et al.* (2010), torna-se necessário definir dois tipos: a anotação lexicográfica e a de texto corrido. A lexicográfica é aquela que se concentra numa unidade lexical em particular e busca sentenças que a contenham em diferentes textos do *corpus*, selecionando, posteriormente, algumas para serem anotadas. Essa é a forma de anotação mais comum na *FrameNet*. Já na anotação de texto corrido, o texto inteiro é anotado para todas as ULs que ele contém. Nos dois tipos de anotação, declara-se cada palavra de uma sentença como alvo, seleciona-se o *frame* em relação ao qual o alvo deve ser anotado e o elemento de *frame* apropriado depois se anota os constituintes relevantes, que fazem parte da valência desse uso. A anotação desses elementos é feita, graficamente, em, pelo menos, três camadas que representam a função semântica, isto é, o Elemento de *frame* que corresponde (por exemplo, Comunicador) a sua descrição sintática, isto é, o Tipo de Sintagma (por exemplo, SN), e a sua Função Gramatical (por exemplo, Externo).

Segundo Salomão (2009), as categorias que resultam da anotação provêm da definição do *frame*. Assim sendo, os Elementos de *frame* (EF) são específicos de cada *frame* e as categorias sintáticas, Tipo de Sintagma (TS) e Função Gramatical (FG) fazem parte da análise gramatical, que, apesar de não declarada formalmente, é seguida pelo projeto, ou seja, os pressupostos das teorias sintáticas construcionistas e não derivacionais, como por exemplo, a Teoria da Sintaxe mais simples, HPSG (*Head-Driven Phrase Structure Grammar*) e, em especial, da Gramática das Construções.

Se o tipo semântico básico dos EFs não for largamente constante na sua utilização, é necessário criar um EF distinto. Os EFs são classificados como nucleares, periféricos e extra-temáticos. Segundo Ruppenhofer *et al*. (2010), um elemento nuclear é aquele que instancia um componente conceptual necessário ao *frame*, tornando-o único e diferente dos demais. Por exemplo, no *frame* de Comunicação, o comunicador, o meio, o tópico e a mensagem são nucleares de acordo com a *FrameNet* do inglês. Existem algumas propriedades formais que determinam quais elementos de *frame* são nucleares.

Elementos de *frame* periféricos são aqueles que não introduzem eventos independentes, distintos ou adicionais em relação ao evento principal reportado. São responsáveis por noções como tempo, espaço, modo, meio, grau e não caracterizam um *frame* único. Além disso, podem ser instanciados em qualquer *frame* semântico que seja apropriado.

Elementos de *frame* extra-temáticos são aqueles que combinam muitos outros *frames*, trazendo-os para o seu escopo ou elaborando descrições de participantes ou de lugares. Assim, os EF extra-temáticos não fazem parte do *frame* no qual eles aparecem. São EF de um outro *frame* abstrato que os utiliza e utiliza também o alvo que eles modificam como argumento.

Algumas vezes, apesar de descritos, os EFs não aparecem nas sentenças escolhidas para anotação. No entanto, apesar da sua omissão, ela deve ser indicada já que oferece informação relevante. O EF identificado mostra qual é o papel semântico que o elemento ausente deveria ocupar. Existem três tipos de omissão focadas na instanciação nula do verbo na qual o fenômeno é mais clara: Instanciação Nula Definida (IND), Instanciação Nula Indefinida (INI) e Instanciação Nula Construcional (INC).

2.1 Procedimentos de anotação de sentenças

Para descrever os *frames* da Língua Portuguesa, seguiremos os passos do processo de análise lexical da *FrameNet* (FILLMORE, 2009):

1º Caracterizar o *frame*;

2º Descrever e nomear os elementos de *frame*;

3º Selecionar as unidades lexicais que pertencem ao *frame*;

4º Anotar exemplos de sentenças extraídas de um *corpus* que mostrem as formas como as unidades lexicais do *frame* fornecem informações relevantes a ele linguisticamente;

5º Gerar automaticamente entradas lexicais e descrições de valência que resumem as observações derivadas da análise dessas entradas.

A princípio, a descrição dos *frames* para o português seguirá o modelo dos *frame*s descritos pela *FrameNet* americana. No entanto, torna-se necessário realizar a anotação das sentenças no *corpus* para comprovar a equivalência entre os *frames.*

O projeto *FrameNet* Brasil trabalha com seis *corpora* que também serão usados nessa análise. Três deles (NILC/São Carlos, ANCIB, ECI-BR) fazem parte dos *corpora* disponibilizados pelo site Linguateca (http://www.linguateca.pt/), "um centro de recursos - distribuído - para o processamento computacional da língua portuguesa". Os outros três (Nurc-RJ, Legenda de filmes, Domínio Público) estão disponíveis no site *Sketch Engine* (http://www.sketchengine.co.uk/), um sistema de consulta a *corpus* que incorpora esboços de palavra (*word sketches*) e um resumo do seu comportamento.

Depois de selecionados os *corpora*, foi feita uma busca em cada um deles pelo lema dos verbos escolhidos a fim de acessar todas as suas ocorrências. Caso o resultado ultrapassasse 400 ocorrências por *corpus* (limite definido via procedimentos estatísticos), as sentenças são submetidas ao programa SPSS *Statistics* v 17.0 que seleciona, de modo aleatório, um dado número de ocorrências dentre as amostras fornecidas. Assim, o número máximo de sentenças analisadas por *corpus* foi 400, amostragem tida como suficiente para a obtenção dos dados necessários.

A seguir, essas sentenças foram copiadas para o programa Microsoft Excel para que fosse possível classificá-las, de acordo com o seu sentido: sentido alvo (1), sentido figurativo (2), usos como adjetivo (3), usos como substantivo (4), contexto insuficiente ou ambíguo (5), outros (6). Como nosso estudo se restringe aos verbos cujo significado está associado à definição de cada um dos *frames* a serem estudados, selecionaram-se apenas as sentenças do tipo 1 para anotação.

Segundo Ruppenhofer (2010), a anotação das sentenças é feita em três camadas, podendo chegar até cinco. A primeira camada anota os EFs, a segunda anota a Função Gramatical de cada EF identificado na sentença, e a terceira anota o Tipo Sintagmático deste constituinte. A quarta camada é denominada de acordo com a classe sintática do alvo, no nosso caso Verbo. Ainda é possível a criação de uma quinta camada denominada Outros para elementos que não se enquadrem nas quatro primeiras camadas (Tabela 1).

TABELA 1
Camadas presentes na anotação

Nós comunicamos os alunos disso ...
Camadas
Nós
COMUNICAR
os alunos
disso
EF
Comunicador
Destinatário
Mensagem
FG
Ext
Obj
Obj
TS
SN

Continua

SN SP
Verbo
Outros

As Funções Gramaticais (FG) só são anotadas para os Elementos de *Frame*; as unidades-alvo jamais são etiquetadas em relação às FG que desempenham na sentença que aparecem.

Os Constituintes que aparecem nas posições sintáticas nucleares são anotados com Externo (EXT) e Objeto (OBJ). Os outros constituintes que seguem o núcleo sintático são anotados como Dependentes (Dep), pois sua instanciação é licenciada pelo núcleo. Elementos obrigatórios (complementos) ou opcionais (adjuntos) não precisam ser diferenciados, pois a definição do *frame* já os distingue em EF nucleares e periféricos. Dessa forma, o rótulo Dep aplica-se a todos os modificadores adjetivais e adverbiais e apesar de ser um rótulo estritamente sintático, sua contribuição é de enriquecimento semântico.

Os tipos sintagmáticos descrevem a valência sintática da UL-alvo e, embora não ofereçam uma análise sintática completa da sentença, capturam os requisitos gramaticais essenciais para as ULs-alvo, inclusive os constituintes relativizados ou extrapostos, que estão fora da localidade sintática. Além disso, marcam sintaticamente os EFs nucleares, periféricos e extra-temáticos.

3 *Frame* de comunicação

A tabela 2 mostra a descrição do *Frame* de Comunicação do Português, incluindo sua definição, os elementos nucleares e não-

nucleares e as relações entre *frame*s, conforme o padrão da *FrameNet* americana.

TABELA 2
Descrição do *Frame* de Comunicação do Português

Definição:

Um Comunicador envia uma Mensagem para um Destinatário; o Tópico e o Meio podem também ser expressos. Esse *frame* não inclui especificação do método de comunicação (escrito, discurso, gestual). Os *frame*s que herdam o *frame* geral de Comunicação podem elaborar o Meio de várias maneiras (em Francês, em um programa de rádio, em uma carta) ou o Modo (murmurou, balbuciou). Existem *frame*s que não herdam todas as características desse *frame* ou que alteram alguma de suas partes.

Nucleares (Core):
Comunicador [Communicator]
Tipo Semântico:
Consciente
É uma entidade consciente que usa a linguagem de maneira oral ou escrita para passar uma Mensagem à outra pessoa.

- Maluf **COMUNICA** sua candidatura ao PPR.

Meio [Medium]
O espaço físico ou abstrato no qual a Mensagem é transmitida.

- Por exemplo, como é que você diz quando você vai **COMUNICAR** com alguém no telefone? [INI]

Mensagem [Message]
Tipo Semântico:
Mensagem
Uma proposição ou um conjunto de proposições que o Comunicador quer que o Destinatário acredite ou tenha como certa.

- Com 30 capítulos já gravados, 5 milhões de cruzeiros investidos na produção, um elenco de 30 atores contratados, a 10 dias da estréia o Serviço de Censura e Diversões Públicas da Polícia Federal, **COMUNICAVA** à Globo que a novela não iria ao ar.

Tópico [Topic]
É a entidade a qual a proposição ou as proposições transmitidas estão relacionadas ou falam sobre.

- Para **COMUNICAR** seu projeto às parcelas mais simples da população espanhola, deu às imagens da Constituição de 1812 atributos da iconografia religiosa tradicional. [IND]

Não-nucleares (Non-core):

Destinatário [Addressee]
Tipo Semântico:
Consciente

Aquele que recebe uma Mensagem do Comunicador.

- O profissional que não puder se apresentar ao trabalho o deve **COMUNICAR** à empresa com antecedência para evitar sanções disciplinares [INI]

Quantidade_
de_informação [Amount_
of_information]
A quantidade de informação trocada quando a comunicação ocorre.

Continua

Descrição [Depictive]
Descreve o estado do Comunicador.

- O edital, uma folha de papel timbrado da prefeitura -- **COMUNICA** a suspeita de caso de Aids em nosso município, Sérgio Barbosa – comunica e relaciona medidas de prevenção à Aids.

Duração [Duration]
Tipo Semântico:
Duração

É o tempo durante o qual a comunicação ocorre.

- -- Não, é muito mais curta; **DIZ**-se em cinco minutos. Tirei o relógio para ver a hora exata, e marcar o tempo da narração. Rita começou e acabou em dez minutos. Justamente o dobro. [INC] [INI]

Frequência [Frequency]
Descreve o número de vezes por unidade de tempo no qual o Comunicador transmite uma Mensagem.

- As conclusões do Provedor serão sempre **COMUNICADAS** aos órgãos ou agentes afetados e, se tiverem origem em uma queixa apresentada, aos reclamantes. [INC]

Modo [Manner]
Tipo Semântico: Modo
O Modo no qual a comunicação ocorre.

- O vazamento atropelou a estratégia do governo, impedindo que ela fosse oficialmente **COMUNICADA** de que poderia continuar ministra, mas de outra pasta, a da Agricultura de que poderia continuar ministra, mas de outra pasta, a da Agricultura

Canal [Means]
A ação que o Comunicador realiza para se comunicar de uma maneira particular descrita pelo alvo.

- o que tenho que comunicar, porque eu acho que você, é muito mais fácil você falar o que você, é importante mesmo, e, e que você se **COMUNICA**, realmente, com uma pessoa falando pessoalmente. [INI]

Lugar [Place]
Tipo Semântico:
Relação_Locativa
É o Lugar onde a comunicação acontece.

- O boato sobre a intenção de Lula em unificar as igrejas evangélicas foi **COMUNICADO** ao comitê central da campanha por dirigentes do PT em vários Estados.

Finalidade [Purpose]
Tipo Semântico:
Estado_de_Coisas
A Finalidade pela qual o Comunicador se comunica.

- Quando se comprove a queixa foi realizada com má-fé o Provedor de Justiça **COMUNICARÁ** o fato ao Ministério Público competente, para o início de um procedimento penal de acordo com a lei.

Tempo [Time]
Tipo Semântico: Tempo
O Tempo no qual a comunicação acontece.

- Imediatamente, os policiais **COMUNICARAM** a ocorrência à 21ª Delegacia Policial.

Razão [Reason]
Tipo Semântico:
Estado das coisas

Esse EF identifica a Razão para usar o meio de comunicação.

- Amorim disse que não **COMUNICOU** a Justiça brasileira sobre a denúncia que recebeu em novembro pois recebi informações sem provas [IND]

Core Set
{Mensagem, Tópico}, {Comunicador, Meio}

4 O *frame* de meios de comunicação

A tabela 3 mostra a descrição do *Frame* de Meios de Comunicação do Português, incluindo sua definição, os elementos nucleares e não-nucleares e as relações entre *frame*s, conforme o padrão da *FrameNet* americana.

TABELA 3
Descrição do *frame* de Meios de Comunicação do Português.

Definição:
Esse *frame* trata da comunicação do Comunicador com outros com o apoio de um Meio de comunicação como o telefone.
Nucleares (Core):
Destinatário [Addressee]
Tipo Semântico:
Consciente
Destinatário é a pessoa para quem a Mensagem é transmitida. Quando esse EF está presente, ele, geralmente, aparece em um sintagma preposicional introduzido por para ou como um objeto direto ou como o primeiro objeto em uma construção de dois objetos.

- O líder do governo na Câmara, Luis Carlos Santos, **TELEFONOU** para o presidente da Casa, Luis Eduardo Magalhães, para dar notícias sobre o andamento das negociações do acordo da Previdência.

Comunicador [Speaker]
Tipo Semântico:
Consciente
Esse EF é a pessoa que usa o Meio de Comunicação para enviar uma Mensagem para o Destinatário. O Comunicador é normalmente expresso como um Argumento Externo do verbo alvo ou como um modificador Genitivo do nome.

- Romário **TELEFONOU** ontem para o Canecão pedindo ingressos para o show Fina estampa, de Caetano Veloso.

Meio [Means]
Tipo Semântico: Estado de Coisas
Esse EF se refere ao instrumento de comunicação usado pelo Comunicador (e Destinatário).

- Ninguém **TELEFONOU** para mim para verificar.

Mensagem [Message]
Tipo Semântico:
Mensagem
Esse é o EF que identifica o conteúdo que o Comunicador está comunicando ao Destinatário. Pode ser expresso como uma oração ou um sintagma nominal.

- Por volta das 17h, ela **TELEFONOU** para o setor internacional da Sony no Rio avisando da mudança de planos .

Tópico [Topic]
É o assunto do qual trata a Mensagem. É tipicamente expresso como um complemento SP iniciado por sobre.

- De Buenos Aires **TELEGRAFAM** sobre a revolução do Paraguai: Partiu para a Vila del Pilar uma comissão de comerciantes, a fim de negociar a paz. [IND] [INI]

Continua

Não-nucleares (Non-core)

Descrição [Depective]

O estado do Comunicador ou da Mensagem durante a comunicação.

- Um diretor da Levi Strauss fabricante das famosas calças Levis **TELEFONOU** para o seu gerente em Bangladesh, louco da vida, por saber que havia cerca de 40 menores trabalhando na fábrica local.

Modo [Manner]

Tipo Semântico: Modo

Esse EF é para expressões que elaboram a maneira na qual uma ação é executada.

- Não há buracos nas ruas, me deixaram **TELEFONAR** de graça e um motorista de táxi não cobrou a corrida porque eu não tinha troco. [INI] [INI]

Razão [Reason]

Tipo Semântico:

Estado das coisas

Esse EF identifica a Razão para usar o meio de comunicação.

Lugar [Place]

Tipo Semântico Relação_locativa

O lugar onde a comunicação acontece.

- perto. - Podemos **TELEFONAR** de uma fazenda. - Estamos ferrados agora. - Vamos para alguma fazenda tentar TELEFONAR. - Por hoje é só Freddy. - Você sabe que isso não ajudará em nada. - Você deve ter calma, como [IND] [INI] [INI]

Frequência [Frequency]

Descreve o número de vezes por unidade de tempo no qual o Comunicador transmite uma Mensagem.

- **TELEFONEI** várias vezes para marcar também a minha festa de formatura, e o dono nunca estava. [IND] [IND] [IND]

Tempo [Time]

Tipo Semântico:

Tempo

O tempo no qual o evento acontece.

- Tinha **TELEFONADO** antes para saber. [IND] [INI] [IND]

Finalidade [Purpose]

Tipo Semântico:

Estado_de_Coisas

A Finalidade pela qual o Comunicador se comunica.

- A (s) pessoa (s) interessada (s) no seu anúncio anotará o código do leitor e **TELEFONARÁ** para 900.0220 para deixar gravada uma mensagem.

Core Set

{Mensagem, Tópico}

5 O *frame* de modos de comunicação do português

A tabela 4 mostra a descrição do *Frame* de Modos de Comunicação do Português, incluindo sua definição, os elementos nucleares e não-nucleares e as relações entre *frame*s, conforme o padrão da *FrameNet* americana.

TABELA 4
Descrição do *frame* de Modos de Comunicação do Português

Definição:
As palavras desse *frame* descrevem Modos de comunicação verbal. Todas podem aparecer com citações.

Nucleares (Core):
Tópico [Topic]
Tópico é o assunto da Mensagem comunicada. É normalmente expresso como um Complemento SP iniciado por sobre e, nesse *frame*, é frequentemente precedido por um substantivo quantificador que faz referência a mensagem.

- Se **COCHICHARÁ** sobre distribuição de cargos e dólares aliciando votos. [INC] [INI]

Destinatário [Addressee]
Tipo Semântico:
Consciente
Destinatário é a pessoa com quem o Comunicador está se comunicando. Quando expresso, aparece como um complemento SP.

- Regina Priolli recém-chegado de Palm Beach, onde tem uma casa só para receber os amigos -- **COCHICHOU** para a amiga Celina Amaral Peixoto: Você namoraria o Caetano? "

Mensagem [Message]

Tipo Semântico:
Mensagem
Mensagem é o conteúdo que é comunicado pelo Comunicador. A mensagem pode ser uma citação direta, uma oração complementar finita ou um objeto SN.

- Quando ia na casa da minha avó, no Rio Comprido, era muito mal vista pois ia de short ou calça comprida, e as mulheres **COCHICHAVAM**: Essa é mulher-homem. [INI]

Comunicador [Speaker]
Tipo Semântico:
Consciente
Comunicador é uma entidade consciente que produz a Mensagem ou comunica sobre um Tópico. É expresso como Argumento Externo do verbo.

- O Itamar ainda **COCHICHOU** em seu ouvido: [INI] [INI]

Não-nucleares (Non-core):
Grau [Degree]
Esse EF identifica o grau em que um evento ocorre.

Descrição [Depective]
Descrição é usado para qualquer sintagma que descreve um participante numa ação.

Continua

Duração [Duration]
Tipo Semântico: Duração
A quantidade de tempo que a comunicação gasta.

Modo [Manner]
Tipo Semântico: Modo
Modo que acontece a comunicação.

- não é a uma penitenciária. - E nem para suas casas. - O que está acontecendo aqui? O que **COCHICHANDO** como dois ratos estão? - Estamos consertando as lâmpadas que estão queimadas. – Tem várias que parecem [IND] [IND]

Canal [Means] **Tipo Semântico:** Ação_humana
A ação que o Comunicador realiza para se comunicar de uma maneira particular descrita pelo alvo.

Meio [Medium]
Esse elemento de *frame* expressa o meio de comunicação, como a linguagem usada ou tipo específico de texto onde a mensagem aparece. É normalmente expresso com um complemento SP iniciado por por ou em.

- Não fala – murmura, **COCHICHA**, em gíria arrevezada. E maltrapilha e zambra, arrasta andrajos e oscila. A praia de Santo Cristo tem o aspecto sadio de uma varina, criada livremente [IND] [INI] [INI]

Lugar [Place]
Tipo Semântico Relação_locativa
O lugar onde a comunicação acontece.

- não, certo? – Fizeram um grande trabalho lá. - Levantem essas cabeças! - Fleury? - Me diga o que **COCHICHOU** pra Janet no escritório, - que a fez parar de chorar pelo o Frank... - por tudo isso que estamos [IND]

Resultado [Result]
Esse EF identifica o resultado de um evento.

Reversivo [Reversive]
Esse EF indica que o ato de comunicação denotado pelo alvo é resposta a um ato de comunicação anterior no qual os papéis do Comunicador e do Destinatário eram invertidos. Ele está conceitualmente relacionado ao *Frame* Resposta_Comunicação no qual o alvo por ele mesmo denota a natureza de resposta do ato.

Tempo [Time]
Tipo Semântico:
Tempo
O tempo no qual o evento acontece.

- Não precisa ficar envergonhada por ter me dito coisas tão adoráveis. - Quer saber o que o Capitão **COCHICHOU** comigo naquele dia? - O que ele me disse, foi que meu verdadeiro amor, - estava bem na frente dos meus

Frequência [Frequency]
Descreve o número de vezes por unidade de tempo no qual o Comunicador transmite uma Mensagem.

- Quem? - Uma que sentava perto de vc na escola primária - Fong. - Fong Chu? - A garota que sempre **COCHICHAVA** com você? - Sim. - Ela não tinha ido estudar na américa? - Ela está de volta? - A família dela [INI]

Finalidade [Purpose]
Tipo Semântico:
Estado_de_Coisas
A Finalidade pela Qual o Comunicador se comunica.

- por ouro e suspendeu essa conversão pura e simplesmente, no grito, eh, no grito, que o outro lado também soube gritar, mas soube **GRITAR** procurando ajeitar o negócio. E até agora eles estão procurando ajeitar, né? Eh, valorizam o marco, valorizam o iene, valorizam [IND] [INI] [INI]

Core Set
{Mensagem, Tópico}

6 Adaptações sugeridas

A princípio, a descrição do *frame* de Comunicação do Português foi estabelecida com base na da *FrameNet* do inglês. No entanto, quando iniciamos as anotações nos *corpora*, percebeu-se que seria necessário fazer algumas adaptações.

Primeiramente, notamos que o EF Destinatário, classificado como elemento periférico no *frame* da *FrameNet* americana, apresentou-se como um elemento essencial no *frame* da Língua Portuguesa, o que o caracterizaria como elemento nuclear. Vejamos o exemplo abaixo:

- Por exemplo, como é que você diz quando você vai se **COMUNICAR** com alguém no telefone? [INI]

Além disso, podemos levar ainda em consideração o fato de o Destinatário ser EF nuclear nos outros dois *frame*s descritos para o inglês analisados, Meios de Comunicação e Modos de comunicação, cujos verbos apresentam um padrão de comportamento bem parecido com o do *frame* de comunicação. Nos exemplos que se seguem, pode-se perceber a presença desse EF em sentenças desses três *frames*.

- On one memorable occasion , she thinks to **COMMUNICATE** her feelings about Catholic beliefs to some of her older pupils . *(Numa ocasião memorável, ela pensa em comunicar seus sentimentos em relação as crenças católicas à alguns do seus alunos)*

Um outro fato relevante é que, para o espanhol, por ser uma língua neolatina como o português, esse EF também é considerado nuclear. Vejamos um exemplo apresentado por eles:

- Sosa **COMUNICÓ** al público que lo escuchaba el júbilo que sentía y su orgullo de ser dominicano , dando repetidas veces gracias a su madre y exhortando a la juventud de su pueblo a perseverar en el trabajo para cosechar frutos futuros

Por fim, a própria *FrameNet* americana apresenta um ponto a favor dessa análise quando define que o *Frame* de Modo de Comunicação herda do *frame* de Comunicação, dessa forma todos os elementos que são nucleares para o pai deveriam ser nucleares para o filho, o que não acontece no caso do EF Destinatário.

Assim sendo, nossa proposta neste trabalho, é que sigamos o modelo espanhol, acrescentando o Destinatário como elemento de

frame nuclear, mas mantendo sua equivalência com o *frame* original da *FrameNet* americana.

Já nos *frame*s de Meios de Comunicação e Modos de Comunicação, foram acrescentados alguns elementos periféricos encontradas na anotação das sentenças dos *corpora*: no caso do primeiro, lugar, frequência, finalidade e tempo; e, para o segundo, frequência e finalidade. No *frame* de Comunicação, também foi acrescentado um elemento periférico: razão. No entanto, como são elementos extra-temáticos, não interferem na equivalência entre *frame*s.

Uma outra sugestão considerada pertinente é uma pequena mudança na definição do EF Comunicador dos três *frame*s. Retomando a definições dos três *frame*s, temos no *frame* de Comunicação a seguinte definição: Comunicador "é uma entidade consciente que usa a linguagem de maneira oral ou escrita para passar uma Mensagem à outra pessoa". No *frame* de Meios de Comunicação, temos: Comunicador "é a pessoa que usa o Meio de Comunicação para enviar uma Mensagem para o Destinatário". E no de Modos de Comunicação: Comunicador "é uma entidade consciente que produz a mensagem ou comunica sobre um tópico."

Percebemos que o Comunicador é sempre uma pessoa ou entidade consciente. Acontece que, na Língua Portuguesa, é comum usarmos uma metonímia, empregando uma instituição no lugar da pessoa, o que não está formalizado na *FrameNet* americana. Veja o exemplo abaixo:

- E depois de tantas negativas, a Mercedes **COMUNICOU** ontem seu casamento com a McLaren.

Apesar de sabermos que a entidade que aparece como Comunicador na sentença, na verdade, representa uma pessoa ou um grupo de pessoas que está se comunicando, é preciso deixar isso claro na definição do *frame*. Assim, propomos as seguintes definições:

Frame de Comunicação: Comunicador: é uma entidade ou pessoa que usa a linguagem de maneira oral ou escrita para transmitir uma Mensagem a outra entidade ou pessoa.

Frame de Meio de Comunicação: Comunicador: é uma entidade ou pessoa que usa o Meio de Comunicação para enviar uma Mensagem para o Destinatário.

Frame de Modos de Comunicação: Comunicador é uma entidade ou pessoa que produz a mensagem ou comunica sobre um tópico.

Um outro fator relevante é em relação ao *CoreSet*. No *frame* de Comunicação da *FrameNet* americana, temos dois *coresets*: {Mensagem, Tópico},{Comunicador, Meio}. No entanto, em seu *frame* de Meios de Comunicação, o *coreset* é formado somente por {Mensagem, Tópico}. No entanto, no processo de anotação, percebemos que o *coreset* {Comunicador, Meio} também poderia estar presente nesse *frame*, representando uma outra metonímia, o meio pela pessoa. Como a *FrameNet* se baseia na análise de *corpus* e como essa figura é presença constante na Língua Portuguesa, assim como outras, é necessário também incluí-las na análise. Vejamos o exemplo abaixo:

- As rádios começaram a **VEICULAR** insistentemente a notícia da ocupação da PM. [INI] [INI]

Existe, ainda, uma particularidade em relação às unidades lexicais dos *frames* de Meios de Comunicação e Modos de Comunicação. Enquanto a *FrameNet* americana lista várias ULs para esses *frames*, na Língua Portuguesa, elas aparecem em menor número. Percebe-se assim que, na Língua Inglesa, é comum incorporar meio e maneira ao sentido do verbo, enquanto na Língua Portuguesa, apesar de existirem casos, eles não são tão comuns como no inglês. Isso se deve de que, no processo de formação de palavras, enquanto as línguas germânicas lexicarizaram movimento e modo, as línguas românicas lexicalizaram movimento e direção. Veja o exemplo de incorporação para o *frame* de Meios de Comunicação:

- She let Herbert know, and he **CABLED** her to cancel all other arrangements and catch the next ship back .DNI *(Ela contou a Herbert, e ele enviou um telegrama a ela para cancelar todos os outros compromissos e pegar o próximo navio de volta.)*

Além disso, no inglês, é possível criar verbos a partir de nomes sem que para isso aconteça alteração morfológica. Já no Português, esse processo implica uma modificação na palavra, como é o caso da derivação parassintética (engarrafar) e da sufixação (telefonar). É comum também o uso de verbos suporte, assim, não existem expressões monolexêmicas correspondentes, mas existem as

polilexêmicas, como por exemplo, "enviar um email", "enviar um telegrama", "comunicar por rádio" entre outras. O mesmo acontece com alguns verbos do *frame* Modos de Comunicação. Veja o exemplo:

- ` I knew it was n't safe , " she **JABBERED** . *(Eu sabia que não era seguro, ela falou rápido e animadamente)*

Existem ULs que representam uma forma tão específica de se comunicar que chega a dificultar a tradução, como é o caso do verbo "*lisp*", que representa o ato de comunicação no qual o comunicador se comunica pronunciando os sons ´s´ e ´z´ como se fossem ´th´.

Assim, para estudarmos esses casos no português, teríamos que procurar nos *corpora* expressões como "enviar um telegrama", no caso de Meios de Comunicação. Já para o Modo de Comunicação, seria ainda mais complicado, já que as expressões não se reduziriam a uma expressão, mas a uma explicação da maneira pela qual o comunicador está se comunicando. Dessa forma, optamos por eleger ULs que tinham sentidos equivalentes.

Percebe-se assim que, apesar de algumas alterações pontuais em relação a definição dos *frames* e a seus elementos nucleares e periféricos, os *frames* analisados são bem parecidos no português e no inglês. Entretanto, no que tange às realizações sintáticas, as Unidades Lexicais do inglês e do português se apresentam de maneira bem distinta.

Apesar de em algumas situações termos um EF com a Função Gramatical Dep no inglês e Obj em português, mesmo assim, marcamos essas ocorrências como semelhantes, visto que acreditamos que essa diferença esteja relacionada à um ponto de vista distinto: enquanto que, no português, consideramos que esses elementos ocupam uma posição sintática nuclear na frase e, por isso, os classificamos como Obj, no inglês, eles são considerados como modificadores e, por isso, são classificados como Dep. Isso se deve ao fato da *FrameNet* americana só considerar como elemento nuclear sujeito e objeto direto. No entanto, acreditamos que os objetos indiretos também podem ser considerados como elementos nucleares no Português, assim, ao invés de classificá-los com Dep, os classificamos como Obj.

No Brasil, ainda não foi criado um *corpus* homogêneo do Português, como é o caso do *British National Corpus* utilizado pela

FrameNet americana, por isso, para fazermos essa análise, utilizamos vários *corpora* criado a partir de fontes bem distintas. Assim, devido a essa falta de homogeneidade dos nossos *corpora,* ou seja, o fato de não termos quantidade semelhante de textos representativos dos vários gêneros e estilos, sabemos que não é possível fazer uma análise definitiva em relação à frequência dos padrões encontrados. No entanto, consideramos pertinente fazer uma comparação das tabelas de padrões das ULs de cada *frame*, a fim de criarmos uma tabela única com os padrões que tiveram uma representação mais significativa por *frame*.

Como nesse momento da análise, tínhamos como objetivo chegar a padrões o mais genérico possível, os tipos sintagmáticos oracionais encontrados (Sfin, Sinf, Sinterrog, Sse, Sger, Srel, Sub, CIT) foram agrupados como SS. Na tabela 5, temos o resultado da comparação dos padrões das cinco ULs do *frame* de Comunicação, que se mostraram bem semelhantes.

TABELA 5
Padrões recorrentes nas ULs do *Frame* de Comunicação

Comunicador	**UNIDADE LEXICAL**	Mensagem
Arg Ext SN		INI
Arg Ext SN		IND
Arg Ext SN		Obj SN
Arg Ext SN		Obj SS
Arg Ext SN		Dep SP
Mensagem	**UNIDADE LEXICAL**	**Comunicador**
Obj SN		INC
Obj SS		IND
Obj SN		IND
Obj SS		Arg Ext SN

UNIDADE LEXICAL	Comunicador	Mensagem
	IND	INI
	IND	IND
	Arg Ext SN	Obj SS
UNIDADE LEXICAL	Mensagem	Comunicador
	Obj SS	IND
	Obj SN	IND

A tabela 6 é o resultado da comparação dos padrões das três ULs do *frame* de Modos de Comunicação. Foram encontrados vários padrões recorrentes nas tabelas analisadas, indicando que esses verbos também têm um comportamento bem parecido.

TABELA 6
Padrões recorrentes nas ULs do *Frame* de Modos de Comunicação

Padrões			
Comunicador Arg Ext SN Arg Ext SN Arg Ext SN Arg Ext SN Arg Ext SN	UNIDADE LEXICAL	Mensagem Obj SN INI Obj SS Obj SS Obj SN	Destinatário INI INI INI IND IND
Comunicador Arg Ext SN Arg Ext SN	UNIDADE LEXICAL	Destinatário Dep SP IND	Mensagem Obj SS INI
Mensagem Obj SS Obj SS Obj SS Obj SS	UNIDADE LEXICAL	Comunicador Arg Ext SN Arg Ext SN IND IND	Destinatário IND INI IND INI
Destinatário Obj SN Obj SN	UNIDADE LEXICAL	Comunicador Arg Ext SN IND	Mensagem INI INI
UNIDADE LEXICAL	Mensagem Obj SN Obj SS Obj SS Obj SN	Comunicador IND IND IND IND	Destinatário INI IND INI IND
UNIDADE LEXICAL	Comunicador IND IND	Mensagem INI INI	Destinatário INI IND
UNIDADE LEXICAL	Destinatário Obj SN	Comunicador IND	Mensagem INI
UNIDADE LEXICAL	Tópico Dep SP	Comunicador IND	Destinatário INI
Mensagem Obj SS	Comunicador Arg Ext SN	UNIDADE LEXICAL	Destinatário INI

Por fim, temos o resultado da comparação da tabela de padrões das ULs do *frame* de Meios de Comunicação. Essas ULs foram as que apresentaram padrões mais distintos, sendo mais difícil encontrar padrões em comum. Apesar de as ULs "telegrafar" e "telefonar" terem comportamento semelhante, os padrões da UL "veicular" são bem discrepantes (tabela 7).

TABELA 7
Padrões recorrentes nas ULs do *Frame* de Meios de Comunicação

Padrões				
Comunicador Arg Ext SN Arg Ext SN Arg Ext SN	**UNIDADE LEXICAL**	Destinatário Dep SP Dep SP INI	Mensagem INI Obj SN INI	
Mensagem Obj SN	**UNIDADE LEXICAL**	Comunicador INC	Destinatário INI	
Destinatário Dep SP	**UNIDADE LEXICAL**	Mensagem IND	Comunicador IND	
Comunicador Arg Ext SN	Destinatário Dep SP	**UNIDADE LEXICAL**	Mensagem INI	
Comunicador Arg Ext SN	Comunicador Arg Ext SN	**UNIDADE LEXICAL**	Mensagem Obj SN	Destinatário INI
UNIDADE LEXICAL	Destinatário Dep SP Dep SP	Mensagem INI Obj SS	Comunicador IND IND	
UNIDADE LEXICAL	Mensagem Obj SN INI INI INI	Comunicador IND IND IND INC	Destinatário INI INI IND INI	

Acreditamos que quando os projetos que estão sendo desenvolvidos para as diferentes línguas estiverem tão avançados quanto o projeto americano e quando foram definidas as metodologias para o trabalho interlinguístico, a comparação entre os *frames* permitirá um conhecimento mais aprofundado das semelhanças e diferenças entre elas, além de colaborar infinitamente para a tradução automática e para a busca semântica.

Referências

FILLMORE, Charles J. *Frame* semantics. In: The Linguistic Society of Korea. (Ed.). *Linguistics in the morning calm*. Seoul: Hanshin, 1982. p. 111-137.

FILLMORE, Charles J.; ATKINS, B. T. Towards a *Frame*-based organization of the lexicon: the semantics of RISK and its neighbors. In: LEHER, Adrienne; KITTAY, Eva. (Ed.). *Frames, fields, and contrasts*: new essays in semantics and lexical organization. Hillsdale: Lawrence Erlbaum, 1992. p. 75-102.

FILLMORE, C. J.; BAKER, C. A *frame*s approach to semantic analysis. In: HEINE, B.; NARROG, H. (Ed.). *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, 2009. p. 313-339.

GAWRON, Jean Mark. *Frame semantics.* Ms. Stanford University. 2008.

PETRUCK, Miriam R. L. *Frame* Semantics. In: VERSCHUEREN, Jef; Ã–STMAN, Jan-Ola; BLOMMAERT, Jan; BULCAEN, Chris. (Ed.). *Handbook of pragmatics.* Philadelphia: John Benjamins, 1996.

RUPPENHOFER, Josef; ELLSWORTH, Michael; PETRUCK, Miriam R. L.; JOHNSON, Christopher R.; SCHEFFCZYK, Jan. *FrameNet II*: extended theory and practice. [on line] 2010. [citado em 06 01 12] Disponível em: http:// www.framenet.icsi.berkeley.edu.

SALOMÃO, Maria M. M. FrameNet Brasil: um trabalho em progresso. *Calidoscópio*, v. 7.3, 2009.

A construção superlativa de expressão corporal: uma análise baseada em *corpora*[1]

Igor de Oliveira Costa[2]
Neusa Salim Miranda[3]

RESUMO: Este trabalho focaliza a dimensão do estudo da Construção Superlativa de Expressão Corporal ("[...] solteirona e toda virgem, ignorava machezas, quase **morreu de vergonha** numa tarde de conversas"; "Padre Dito quase **estourou de rir** [...]"; "O Lúcio **rolou de rir** com a explicação, e como consequência acabou virando a vítima e a cobaia do seminário.") que envolve mais diretamente a utilização de *corpus*. O viés teórico pelo qual é visto o objeto é a Linguística Cognitiva e a Gramática das Construções Cognitiva. O *corpus* que subsidiou a pesquisa é o Corpus do Português (http://www.corpusdoportugues.org/), composto por quarenta e cinco milhões de palavras, distribuídos em cinquenta e sete mil textos dos séculos XIV-XX. Os resultados apontam, dentre outras coisas, para a vantagem em se adotar uma abordagem baseada em corpus na investigação de construções de uma língua, permitindo acesso a um entendimento da produtividade e convencionalização da construção em determinada língua.

PALAVRAS-CHAVE: Linguística Cognitiva. Gramática das Construções Cognitiva. Abordagem baseada em *corpus*. Intensidade. Construções Superlativas.

[1] Este trabalho discute, de maneira sintética, uma das dimensões do estudo da Construção Superlativa de Expressão Corporal, analisada de maneira ampla na dissertação de mestrado de Igor de Oliveira Costa, nomeada "A Construção Superlativa de Expressão Corporal: uma abordagem construcionista", orientada por Neusa Salim Miranda no PPG Linguística/UFJF e defendida em agosto de 2010.

[2] Doutorando em Linguística pela Universidade Federal de Juiz de Fora, bolsista de doutorado da mesma universidade. E-mail: igorsabo@yahoo.com.br.

[3] Doutora em Educação pela Universidade Federal de Minas Gerais, professora da FALE e do PPG Linguística/ Universidade Federal de Juiz de Fora. E-mail: neusasalim@oi.com.br.

ABSTRACT: This work focuses on the corpus dimension of the Superlative Construction of Body Expression ("[...] solteirona e toda virgem, ignorava machezas, quase **morreu de vergonha** numa tarde de conversas"; "Padre Dito quase **estourou de rir** [...]"; "O Lúcio **rolou de rir** com a explicação, e como consequência acabou virando a vítima e a cobaia do seminário."). The theoretical approach involves the Cognitive Linguistics and the Cognitive Construction Grammar. The corpus used is the Corpus do Português (http://www.corpusdoportugues.org/), composed of forty-five million words of fifty-seven thousand texts of the XIV-XX centuries. The results points, among other things, to the advantage in adopting a corpus based approach on the constructions' investigation, once it offers access to the comprehension of the construction's productivity and conventionalization in a language.

KEYWORDS: Cognitive Linguistics. Cognitive Construction Grammar. Corpus based approach. Intensity. Superlative Constructions.

1 Introdução

A noção de grau é algo muito caro à gramática das línguas. É através de construções escalares, que denotam grau, que os falantes/escritores de uma língua podem aproximar aquilo que falam/escrevem daquilo que viram, experienciaram ou creem experienciar, dentre outras coisas.

Muitas são as estruturas na Língua Portuguesa (assim como em outras línguas) que servem a esse propósito, de intensificar um enunciado. Mas, na contra mão do uso que os falantes/escritores fazem de tais construções, a tradição gramatical e mesmo a tradição linguística, pouco, ou quase nada, dedicaram-se ao estudo de tal fenômeno. Alguns exemplos de construções modificadoras de grau que tiveram e têm espaço, por exemplo, nas gramáticas normativas são: Construções Comparativas ("Ele é **tão** rápido **quanto** o Bolt"/"Eu escrevo **melhor/pior do que** ele"), Construções com Advérbios de Intensidade ("Maria Fernanda Cândido é perfeita **demais**"), expressões pleonásticas ("Que jogada **linda, linda, linda**!").

Em vista de tal lacuna, o presente trabalho, juntamente com outros, visa ampliar o estudo das manifestações do grau na Língua Portuguesa, como forma de contribuir para uma descrição mais ampla

da língua. O objeto aqui investigado é a Construção Superlativa de Expressão Corporal (doravante, SEC):

(1) 19:Fic:Br:Cony:Piano Enquanto o sábado não chegasse, ele podia se **fartar de ouvir** todos os discos que quisesse [...]

(2) 19Or:Br:Intrv:ISP [...] o meu clown não consegue cruzar os braços. A platéia **morre de rir** do que é, na verdade, uma tragédia para o meu personagem.

(3) 19:Fic:Br:Garcia:Silencio [...] queria era apenas assustar, podemos telefonar para ele e dizer que eu **estou me borrando de medo**.

Por se tratar de uma pesquisa bastante ampla (que, além da descrição formal e semântico-pragmática, envolve as motivações conceptuais, as relações de herança nas quais a construção está envolvida, seu processo de gramaticalização, dentre outras questões[4]), neste trabalho, recortamos a parte do estudo da SEC que está mais diretamente relacionado ao uso de *corpora*.

Esta pesquisa vincula-se ao macroprojeto "Construções Superlativas do Português do Brasil: um estudo sobre a semântica de escalas" (MIRANDA, 2008 – CNPq), que, de sua gênese até o momento, elucidou, contando com o estudo da SEC, sete nódulos[5] dessa grande rede de construções, além das outras quatro pesquisas que tangem o tema que ainda estão em andamento.

O trabalho se organiza da seguinte maneira: a primeira seção traz a perspectiva teórica pela qual se entende o objeto; a seção seguinte aborda a metodologia eleita na investigação e o processo de coleta de dados; a seção 3, por sua vez, trará as análises da SEC que envolvem o uso do *corpus*; ao final serão apresentadas as considerações finais, seguidas dos agradecimentos e das referências.

[4] O trabalho empreendido em Costa (2010) cobre a maior parte de tais questões.

[5] Os outros nódulos da grande rede de Construções Superlativas investigados no interior do macroprojeto estão em Sampaio (2007), Carvalho-Miranda (2008), Albergaria (2008), Miranda (2008b), Carrara (2010) e Machado (2011).

2 Referencial teórico

O quadro referencial teórico deste estudo é composto pela Linguística Cognitiva (FAUCONNIER, 1994; FAUCONNIER; TURNER, 2002; FILLMORE, 1982; FILLMORE; ATKINS, 1992; JOHNSON, 1987; LAKOFF, 1987; LAKOFF; JOHNSON, 1980, 1999; MIRANDA, 2002, 2008a, 2008b; SALOMÃO, 1997, 2006; dentre outros) e um de seus modelos de gramática, a Gramática das Construções Cognitiva (GOLDBERG, 1995, 2006; BOAS, no prelo).

O programa cognitivista de investigação da linguagem, surgido ao final da década de setenta do século passado, contrapõe-se fortemente ao modelo gerativista e a modelos de semânticas vericondicionais. De maneira geral, o modelo sociocognitivista entende (1) a linguagem como uma faculdade cognitiva não autônoma, regulada por aparato cognitivo geral; (2) advoga um papel central para processos imaginativos (metáfora, metonímia, mesclagem) na cognição e na linguagem humanas; (3) vê a gramática como conceptualização, como uma forma de perspectivar uma cena humana; e (4) assume que o conhecimento sobre a linguagem emerge de seu uso.

A Gramática das Construções Cognitiva (CCxG) (GOLDBERG, 1995, 2006; BOAS, no prelo), definindo construções como pares de forma-função, confere a tais estruturas o estatuto de unidades básicas da linguagem. Nesse enquadre, a gramática e o léxico se definem como uma rede de construções instituídas pelo uso através da cultura. A descrição de tais estruturas, dessa forma, passa pela definição não apenas de seus padrões formais, mas também de suas dimensões de sentido e seu uso.

Ponto chave para o modelo goldberiano de gramática são as variáveis frequência de *token* e frequência de *types*, responsáveis, respectivamente, pelo entrincheiramento de determinado padrão na mente dos falantes de uma língua e pela convencionalização de uma construção em determinada língua, ou seja, pela capacidade de uma construção se estender a casos novos dentro da língua. Assim, por permitir a verificação de tais dados, o estudo de um objeto como o deste trabalho com base em *corpus* é altamente profícuo e produtivo.

Como modelo de gramática totalmente imersa nos pressupostos da Linguística Cognitiva, a Gramática das Construções oriundas dos trabalhos de Goldberg visa oferecer explicações psicologicamente

plausíveis para a linguagem (CROFT; CRUSE, 2004, p. 272; BOAS, no prelo, p. 12), explorando relações de motivação e herança entre as construções.

3 Metodologia

Em virtude do relevo do uso no modelo teórico-analítico adotado (a Gramática das Construções Cognitiva é um modelo de linguagem baseado no uso, cf. CROFT; CRUSE, 2004, p. 291-327), acolhe-se uma abordagem baseada em *corpus* (ALUÍSIO; ALMEIDA, 2006; GRIES; DIVJAK, 2003; SARDINHA, 2004; STEFANOWITSCH, 2006) na investigação do objeto.

A montagem de um banco de dados envolvendo especificamente casos da SEC constitui o primeiro (e decisivo) passo no estudo de uma construção, pois é uma forma de deixar os dados falarem, e não ficar reféns unicamente de nossas intuições. Por isso, no intuito de sermos fieis a isso, dividiu-se em duas diferentes fases a busca por casos da construção: uma em que utilizamos de diferentes fontes para levantar os mais diferentes *types* da construção e outra em que nos valemos de um *corpus* anotado para o estudo mais sistemático da construção.

Primeira fase: partindo dos resultados de Sampaio (2007) que aponta, no padrão X DE Y, "rir" como elemento Y mais frequente (*chorar de rir*, *fartar-se de rir*, *morrer de rir* etc.), primeiramente investigou-se, em três diferentes bancos de dados de linguagem (*Corpus* do Português, *Corpora* do Projeto VISL e portal Abril.com), a expressão "de rir" como forma de levantar os elementos X de nosso padrão construcional (o anexo I traz os dados obtidos a partir dessa estratégia). A hipótese inicial era de que, partindo de um *type* mais frequente e, por isso, mais convencionalizado, obter-se-ia uma gama ampla e significativa de variáveis combinatórias deste padrão construcional. De fato, nossa hipótese se confirmou. Os seguintes *types* foram levantados a partir das buscas:

TABELA 1
Types da SEC levantados
Fonte: Costa (2010, p.80-81)

	Type da construção **(Y = rir)**	**Corpus do Português** (45 milhões de palavras)	**VISL** (360 milhões de palavras)	**Abril.com** (não disponível)	**Total**
01	ACABAR(-SE) de rir	—	—	09	09
02	BORRAR(-SE) de rir	01	—	—	01
03	CAGAR(-SE) de rir	—	—	01	01
04	CAIR de rir	—	—	01	01
05	CANSAR(-SE) de rir	01	02	—	03
06	CHORAR de rir	01	—	03	04
07	CONTORCER(-SE) de rir	—	01	01	02
08	DOBRAR(-SE) de rir	—	—	03	03
09	ENGASGAR(-SE) de rir	—	01	—	01
10	ESBALDAR(-SE) de rir	—	—	01	01
11	ESBORRACHAR(-SE) de rir	—	—	01	01
12	ESCANGALHAR(-SE) de rir	—	—	09	09
13	ESCRACHAR(-SE) de rir	—	—	01	01
14	ESGANIÇAR(-SE) de rir	—	—	01	01
15	ESPREMER(-SE) de rir	—	01	—	01
16	ESTOURAR(-SE) de rir	01	—	—	01
17	FARTAR(-SE) de rir	10	19	—	29
18	FINAR(-SE) de rir	01	—	—	01
19	MIJAR(-SE) de rir	—	01	01	02
20	MORRER de rir	14	20	185	219
21	NÃO SE AGÜENTAR de rir	—	—	01	01
22	PASSAR MAL de rir	—	—	02	02
23	RACHAR(-SE) de rir	—	—	08	08
24	RASGAR(-SE) de rir	—	—	01	01
25	REBENTAR(-SE) de rir	01	—	—	01
26	ROLAR de rir	—	08	52	60
27	TORCER(-SE) de rir	—	—	01	01
	TOTAL	30	53	282	365

Segunda fase: munidos de tais dados, passamos à busca de cada um dos vinte e sete (27) *types* verbais encontrados em um único *corpus*, o Corpus do Português.

O Corpus do Português conta com quarenta e cinco milhões de palavras distribuídas em torno de cinquenta e sete mil textos, de variados gêneros, englobando tanto o Português de Portugal quanto o Português do Brasil, perpassando os séculos XIV a XX. O *corpus* não conta com textos do século corrente, o que, aliada à natureza mais formal dos textos que o compõem, representam um certo prejuízo para a pesquisa, mas que de forma alguma a invalida ou mesmo a faz menos relevante.

O padrão de busca nessa etapa foi o lexema do verbo acrescido da preposição "de" (e.g.: [fartar] de; [morrer] de), para que pudéssemos abarcar em uma única busca todas as flexões dos verbos. Realizada as buscas, o *corpus* então nos trouxe como resultados todos os casos em que tais verbos precedem à preposição "de". Como nem tudo são ocorrências da construção que investigamos, foi necessário realizar uma "limpeza" manual dos dados, excluindo os casos que envolviam os verbos buscados seguidos da preposição "de" que não eram casos da SEC.

A tabela abaixo resume os dados obtidos nesta etapa. A coluna "Ocorrências da SEC" representa o volume de dados final que serviram à análise da construção.

TABELA 2
Os dados obtidos na segunda fase da pesquisa
Fonte: Costa (2010, p. 84)

	***Type* da SEC**	**Resultados da busca**	**Ocorrências da SEC**	**Produtividade da busca**
01	ACABAR(-SE) de Y	252	08	3,2%
02	BORRAR(-SE) de Y	08	04	50%
03	CAGAR(-SE) de Y	03	02	66,7%
04	CAIR de Y	835	96	11,5%
05	CANSAR(-SE) de Y	437	372	85,1%
06	CHORAR(-SE) de Y	196	112	57,1%
07	CONTORCER(-SE) de Y	06	01	16,7%
08	DOBRAR(-SE) de Y	75	01	1,3%

Continua

09	ENGASGAR(-SE) de Y	—	—	—
10	ESBALDAR(-SE) de Y	—	—	—
11	ESBORRACHAR(-SE) de Y	—	—	—
12	ESCANGALHAR(-SE) de Y	01	01	100%
13	ESCRACHAR(-SE) de Y	—	—	—
14	ESGANIÇAR(-SE) de Y	—	—	—
15	ESPREMER(-SE) de Y	06	—	—
16	ESTOURAR(-SE) de Y	27	17	63%
17	FARTAR(-SE) de Y	401	381	95%
18	FINAR(-SE) de Y	18	05	27,8%
19	MIJAR(-SE) de Y	02	01	50%
20	MORRER de Y	1.486	674	45,4%
21	NÃO SE AGÜENTAR de Y	01	01	100%
22	PASSAR MAL de Y	—	—	—
23	RACHAR(-SE) de Y	18	01	5,6%
24	RASGAR(-SE) de Y	46	05	10,9%
25	REBENTAR(-SE) de Y	52	34	65,4%
26	ROLAR de Y	29	—	—
27	TORCER(-SE) de Y	30	10	33,3%
	TOTAL	3.929	1.726	43,9%

4 Análises

Na tarefa descritivo-explicativa da SEC, alguns achados estão mais fortemente ligados à adoção de *corpus* na pesquisa. Conforme explicitado à introdução, são tais achados que passamos a apresentar.

Em vista dos dados obtidos através do *corpus*, a SEC mostra-se como uma construção bastante produtiva, instanciando 19 diferentes *types*. A construção também pode ser considerada convencionalizada, uma vez que 1.726 ocorrências da construção foram encontradas no Corpus do Português. Esse número corresponde a 43,9% do uso dos 19 verbos seguidos da preposição "de" no *corpus* (3.929).

Há, no entanto, uma variação em relação à convencionalização de cada *type*: apenas "Morrer de Y", "Fartar(-se) de Y", "Cansar(-se) de Y", "Chorar de Y", "Cair de Y" apresentaram um número de ocorrências que pudessem atestar suas convencionalizações, tal como mostra a tabela 3:

TABELA 3
A convencionalização dos *types* da SEC no *Corpus* do Português

	***Types* da Construção SEC**	***Tokens***
01	MORRER de Y	674
02	FARTAR(-SE) de Y	381
03	CANSAR(-SE) de Y	372
04	CHORAR de Y	112
05	CAIR de Y	96
06	REBENTAR(-SE) de Y	34
07	ESTOURAR(-SE) de Y	17
08	TORCER(-SE) de Y	10
09	ACABAR(-SE) de Y	08
10	FINAR(-SE) de Y	05
11	RASGAR(-SE) de Y	05
12	BORRAR(-SE) de Y	04
13	CAGAR(-SE) de Y	02
14	MIJAR(-SE) de Y	01
15	ESCANGALHAR(-SE) de Y	01
16	CONTORCER(-SE) de Y	01
17	DOBRAR(-SE) de Y	01
18	NÃO SE AGÜENTAR de Y	01
19	RACHAR(-SE) de Y	01
	TOTAL	1.726

De acordo com as ocorrências da SEC no *corpus*, foi possível entender com maior precisão a forma da construção:

$$[X_{V}\ \textbf{de}\ Y_{N/V}],$$

em que X é preenchido por verbos que suscitam o domínio conceptual impacto físico ("acabar", "cagar", "cair", "rachar", "rolar") ou impacto fisiológico ("cansar", "mijar", "morrer") e Y é prototipicamente um nome abstrato ou um verbo:

(4) 16:FMMelo:Letters Com as premissas de que haveria de seguir o Conde Ene ao Brasil, **me acabei de destruir**, empenhar e carregar de novas obrigações.

(5) 18:Machado:Memórias Minha irmã Sabina, já então casada com o Cotrim, **andava a cair de fadiga**. Pobre moça! dormia três horas por noite, nada mais.

(6) 18:Azevedo:Japão [...] dragonas de ouro e desses chapéus de pluma que fizeram **rebentar de medo** o Imperador da China nas profundezas empedradas de Pekin.

(7) 18:Álvares:Lira E quando eu **morra de esperar** por ela Deixai que eu durma ali e que descanse

(8) 19N:Pt:Beira Maria do Carmo Borges, a presidente em exercício, **não se cansou de valorizar** esta festa, e tinha razões para isso.

(9) 19Or:Br:Intrv:ISP Aí Cacá fez Ubu, estourou e eu **fiquei morrendo de inveja**.

(10) 19:Fic:Br:Novaes:Mao Foi quando, quase **se mijando de medo**, o moleque o cutucou com a coronha do bacamarte (...)

O fato de o Corpus do Português ser um *corpus* formado por textos de natureza mais formal (cf. seção 3) impediu a postulação generalizações mais amplas sobre o habitat da SEC. Mesmo assim, os dados levantados nos permitiu entender que se trata de uma construção mais pertinente a contextos discursivos em que o falante/escritor possui maior liberdade de expressão subjetiva, uma vez que é especialmente presente em sequências narrativas e em diálogos (nos textos de ficção, 87,2% de sua ocorrência no corpus utilizado) e em trechos de relatos (outros gêneros).

5 Considerações finais

Propusemo-nos aqui a expor a dimensão do estudo da SEC que está ligada à escolha de se utilizar *corpus* na pesquisa. Ao fazer isso, acabamos por apresentar uma forma, pode-se dizer, eficaz de investigar padrões construcionais em uma língua e as vantagens que

uma abordagem baseada em *corpus* pode oferecer a pesquisas que tenha por foco objetos dessa natureza.

Para formar esse quadro, além de uma brevíssima apresentação das teorias que embasam nosso modo de olhar para o objeto, apresentamos o método que nos valemos na investigação e também dos achados possíveis unicamente porque escolhemos trabalhar com *corpus*: a convencionalização e produtividade da SEC na Língua Portuguesa (através do *corpus*, que é entendido como uma amostra representativa da língua, cf. SARDINHA, 2004, p. 22-25), a descrição da forma da construção e os textos que a construção frequenta.

Os resultados apresentados mostram que, de fato, é vantajosa a utilização de *corpora* na investigação da linguagem, não só por oferecer acesso a informações inacessíveis a introspecção do pesquisador, mas também por permitir descrições mais precisas, e reais, de um determinado objeto, já que as informações emergem naturalmente dos dados.

É certo que o uso de *corpus* não garante uma plenitude de análise (no estudo da SEC, por exemplo, não encontramos através da pesquisa em *corpus* de casos que julgávamos frequentes, como "Pirar de rir"), mas, como afirma Fillmore (1992, p. 35 *apud* SARDINHA, 2004, p. 43), "não há nenhum *corpus* que contenha toda a informação que eu quero explorar, [mas] todo *corpus* me ensinou coisas sobre a linguagem que eu não teria descoberto de nenhum outro modo".

Agradecimentos

O macroprojeto "Construções Superlativas do Português do Brasil: um estudo sobre a semântica de escala" recebe o suporte financeiro do CNPq e o projeto que este estudo integra, a investigação da SEC, recebeu apoio financeiro – bolsa de estudos da FAPEMIG.

Referências

ALBERGARIA, G. *Projeção figurativa e expansão categorial do PB*: o caso de um *frame* "animal". 2008. 107 p. Dissertação (Mestrado em Linguística) – Faculdade de Letras/Universidade Federal de Juiz de Fora, Juiz de Fora, 2008.

ALUÍSIO, S. M.; ALMEIDA, G. M. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa lingüística. *Calidoscópio*, São Leopoldo, v. 4, n. 3, p. 155-177, 2006.

BOAS, H. C. Cognitive Construction Grammar. In: TROUSDALE, G.; HOFFMANN, T. (Ed.). *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press: no prelo.

CARRARA, A. C. F. *As Construções Superlativas Causais Nominais – uma abordagem construcionista*. 2010. 150f. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2010.

CARVALHO-MIRANDA, L. C. *As construções concessivas de polaridade negativa no Português do Brasil*. 2008. 160f. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2008.

COSTA, I. O. *A Construção Superlativa de Expressão Corporal*: uma abordagem construcionista. 2010. 142f. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2010.

CROFT, W.; CRUSE, A. *Cognitive Linguistics*. New York: Cambridge University Press, 2004.

FAUCONNIER, G. *Mental Spaces*. New York: Cambridge University Press, 1994.

FAUCONNIER, G.; TURNER, M. *The way we think*: conceptual blending and the mind's hidden complexities. New York: Basic Books, 2002.

FAUCONNIER, G. Frame semantics. In: Linguistic Society of Korea (Ed.). *Linguistics in the Morning Calm*: Selected Papers from SICOL-1981. Seoul, Hanshin Publishing Co., 1982. p. 111-137.

FILLMORE, C.; ATKINS, B. T. Toward a Frame-Based Lexicon: The Semantics of RISK and its Neighbors. In: LEHRER, A.; KITTAY, E. F. (Ed.). *Frames, Fields and Contrasts*: New Essays in Semantic and Lexical Organization. New Jersey: Lawrence Erlbaum, 1992. p. 75-102.

GOLDBERG, A. *Construction*: A construction grammar approach to argument structure. Chicago: The University of Chicago Press, 1995.

GOLDBERG, A. *Constructions at work*: The nature of generalization in language. Oxford: Oxford University Press, 2006.

GRIES, S. T.; DIVJAK, D. Behavioral profiles: A corpus-based approach to cognitive semantic analysis. In: EVANS, V.; POURCEL, S. (Ed.). *New directions in Cognitive Linguistics.* Amsterdam, Philadelphia: John Benjamins, 2003. p. 57-75.

LAKOFF, G. *Women, Fire and Dangerous Things*: What categories reveal about the mind. Chicago: The University of Chicago Press, 1987.

LAKOFF, G.; JOHNSON, M. *Metáforas da vida cotidiana*. Trad. Mara Sophia Zanotto (Coord.). Campinas: Mercado de Letras; São Paulo: Educ, 2002[1980].

LAKOFF, G. *Philosophy in the Flesh*: The embodied mind and its challenge to western thought. New York: Basic Books, 1999.

MACHADO, P. M. *A Construção Superlativa Sintética de Estados Absolutos com o sufixo "-íssimo"*: um caso de desencontro/mismatch morfológico. 2011. 139f. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2011.

MIRANDA, N. S. O caráter partilhado da construção da significação. *Veredas*, Juiz de Fora, v. 5, n. 2, p. 57-81, 2002.

MIRANDA, N. S. *Construções Superlativas no Português do Brasil*: um estudo sobre a semântica de escala. Projeto de pesquisa do Programa de Pós-Graduação em Letras – Mestrado em Linguística; GP "Gramática e Cognição", CNPq, Universidade Federal de Juiz de Fora, 2008a.

MIRANDA, N. S. *Gramaticalização e gramática das construções*: algumas convergências. Um estudo de caso: as construções negativas superlativas de IPN. 2008. 110 f. Relatório (Pós-doutorado em Linguística) – Centro de Comunicação e Letras, Universidade Presbiteriana Mackenzie, São Paulo, 2008b.

SALOMÃO, M. M. M. Gramática e interação: o enquadre programático da hipótese sócio-cognitiva sobre a linguagem. *Veredas*, Juiz de Fora, v. 1, n. 1, p. 23-39, 1997.

SALOMÃO, M. M. M. *Teorias da Linguagem*: a perspectiva sociocognitiva. Rio de Janeiro, 2006. Disponível em: <http://www.forumdelinguagem.com.br/textos/Texto%20Margarida%20Salom%C3 %A3o.pdf>. Acesso em: 05 out. 2008.

SAMPAIO, T. F. *O uso metafórico do léxico da morte*: uma abordagem sociocognitiva. 2007. 167f. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2007.

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

STEFANOWITSCH, A. Words and their metaphors: A corpus-based approach. In: STEFANOWITSCH, A.; GRIES, S. *Corpus-based Approaches to Metaphor and Metonymy*. Berlin, New York: Mouton de Gruyter, 2006. p. 61-105.

Contribuições metodológicas para o desenvolvimento da plataforma FrameNet Brasil: a descrição de algumas unidades lexicais dos frames `Fechamento` e `Movimento_corporal`[1]

Gabriela da Silva Pires[2]
Margarida Maria Martins Salomão[3]

RESUMO: O presente trabalho é vinculado ao projeto de pesquisa de implantação do Projeto FrameNet Brasil (SALOMÃO, 2009) e tem como objetivo empreender a descrição lexicográfica de sete Unidades Lexicais (ULs) que evocam a cena de abertura no frame `Fechamento`,[4] e três ULs evocadoras do frame `Movimento_corporal`. O respaldo teórico da pesquisa é a Semântica de Frames (FILLMORE, 1982; GAWRON, 2008; PETRUCK, 2008). Este trabalho, assim como a FrameNet como um todo, é ancorado em evidência de *corpus*, razão pela qual a constituição e o adequado tratamento dos *corpora* foram passos importantes no percurso da pesquisa.

PALAVRAS-CHAVE: Semântica de Frames, FrameNet. Frame `Fechamento`. Frame `Movimento_corporal`.

---

[1] Dissertação de mestrado vinculada ao Projeto FrameNet Brasil. Agência de fomento: FAPEMIG. Processo: APQ-01021-08

[2] Mestre em Linguística UFJF (2010) com bolsa pela agência de fomento CAPES. Doutoranda no Programa de Pós-Graduação em Linguística- UFJF. E-mail: gabrielaniger@yahoo.com.br

[3] Professora associada da UFJF. Atua no Programa de Pós-Graduação em Linguística. Líder do Grupo de Pesquisa em Gramática e Cognição. Doutora em Linguística pela Universidade da Califórnia, em Berkeley. E-mail: mm.salomao@uol.com.br

[4] Os frames, por representarem esquemas conceptuais, são distinguidos graficamente. Neste trabalho, estão na fonte `Courier New`.

ABSTRACT: This work is associated to the research project to implement Brazil FrameNet Project (SALOMÃO, 2009a) and it aims at the lexical description of seven Lexical Units (LU) which evoke the opening scene in the `Closure` frame, and three Lexical Units of the `Body_movement` frame. The theoretical references are from the Frames Semantics (FILLMORE, 1982; GAWRON, 2008; PETRUCK, 2008). This work was done based on *corpus* evidences and for this reason the adequate constitution and treatment of *corpora* are important steps of the research.

KEY-WORDS: Frame Semantics. FrameNet, `Closure` Frame, `Body_movement` Frame.

1 Introdução

Este trabalho é um empreendimento em descrição lexicográfica baseada na Semântica de Frames, nos moldes do Projeto FrameNet, desenvolvido para a língua inglesa. Compõe a tríade de trabalhos inaugurais do Projeto FrameNet Brasil (SALOMÃO, 2009), coordenado pela Profa. Dra. Margarida Salomão e desenvolvido na Universidade Federal de Juiz de Fora. O principal objetivo é proceder à análise lexicográfica de um grupo de Unidades Lexicais relacionadas à cena de abertura física, constitutivo do frame `Fechamento` e do frame `Movimento_corporal`.

A descrição lexicográfica aqui empreendida visa a apresentar as combinações sintático-semânticas das chamadas Unidades Lexicais (ULs), que são os termos que evocam determinado frame. Será apresentada descrição lexicográfica de sete ULs do frame `Fechamento`, a saber: *desabotoar*, *desarrolhar*, *desatarraxar*, *destampar*, *abrir_((tampa))*, *levantar_((tampa))* e *tirar_((tampa))*, e de três ULs do frame `Movimento_corporal`: *abrir_((boca))*, *abrir_((mão))*, *abrir_((olho))*. Esse procedimento se deu de acordo com os preceitos da Semântica de Frames (FILLMORE, 1982; GAWRON, 2008; PETRUCK, 2008) e a metodologia de trabalho do Projeto FrameNet (RUPPENHOFER *et al*, 2006).

2 Semântica de frames

A Semântica de Frames se define como uma abordagem que reivindica a continuidade entre língua e experiência e se aplica à

organização de conhecimento. Semântica de Frames como a ilustramos aqui é o modelo fundado pelo linguista Charles Fillmore, fruto do desenvolvimento de diversos estudos iniciados na década de 70, amadurecidos e fortemente difundidos desde o início da década de 80. Esta abordagem empírica tem como noção central o frame, que é um conjunto de conceitos que se encontram estruturados de forma interdependente e são capazes de gerar expectativas. Tal interdependência existe tanto entre os conceitos constitutivos desse conjunto, bem como pelo conjunto em relação às suas partes constitutivas. Conforme diz Fillmore (1982, p. 111), o frame é "qualquer sistema de conceitos relacionados de tal forma que para entender um deles é necessário entender toda a estrutura na qual ele se encaixa" e a introdução de algum desses conceitos faz com que os outros fiquem disponíveis, passíveis de serem acessados.

Um ponto primário da significação para Fillmore, e citado por Gawron (2008), é entender o significado como relativizado ao frame. Considerar o conceito de frame para uma abordagem linguística do significado não é, para Fillmore, vê-lo como um meio extra de organizar conceitos, e sim como fundamental para se repensar as metas de uma semântica linguística. Assim, a Semântica de Frames oferece meios mais atraentes para lidar com questões da significação.

Uma descrição satisfatória de um item lexical deve adotar mais que critérios de traços constitutivos para dar conta de explicar os diversos empregos que um item lexical pode ter. Pensemos no caso de *bachelor*, que requer, para sua compreensão, ser relacionado a questões maiores que simplesmente conceder um homem, adulto, que nunca tenha sido casado. Fillmore (1982) alega que **esse** sentido de *bachelor* (*solteirão*) implica assumir um determinado contexto de expectativas. Assim, para citar alguns exemplos, homens não casados no papel, ou mesmo o Papa, não são colocados no grupo dos solteirões.

Fillmore alega que o objetivo primário da análise do significado é a compreensão e considera a Semântica de Frames como uma semântica da compreensão (*Semantics of Understanding*). Sendo assim, os empenhos do falante (escritor) em atribuir sentido ao seu texto e os empenhos do ouvinte (leitor) em construir sentido para o texto são cruciais. Pensando na interpretação de sentenças, o objetivo na semântica da compreensão é "determinar em qual situação uma sentença se encaixa" (PETRUCK, 2008, p. 3). A importância da Semântica de Frames para a questão da interpretação é apresentada

quando Fillmore (1982, p.117) diz que, por serem estruturas que geram expectativas, os frames atuam para levar à interpretação textual adequada.

Assim, a abordagem da Semântica de Frames oferece uma enorme contribuição no campo da lexicografia, pois busca organizar itens lexicais como estruturas de conhecimento definidas em rede.

3 Metodologia e procedimentos analíticos

3.1 O norteador teórico-metodológico: FrameNet

A FrameNet deve ser entendida como um recurso de descrição lexicográfica baseado na Semântica de Frames. A pesquisa lexicográfica feita pela FrameNet é ancorada em evidências de *corpus*. É um projeto em lexicografia computacional, que extrai de vários *corpora* eletrônicos informações sobre as combinações sintático-semânticas de palavras (Unidades Lexicais) que evocam frames, de tal modo a constituir uma rede semântica baseada em frames. Há procedimentos manuais, envolvendo a anotação de sentenças extraídas de *corpora* e, também, conta-se com o auxílio de *softwares* a partir dos quais são feitos procedimentos automáticos para obter os resultados. Os resultados geram relatórios (em termos de padrões de valência, sumariamento das realizações sintáticas de Elementos de Frame) que estruturam as informações de forma interconectada, formando uma grande rede de significados (FILLMORE, JOHNSON & PETRUCK, 2003).

É necessário fazer uma distinção teórico-metodológica entre a palavra e a Unidade Lexical (UL). A palavra, apesar de ter sua definição bastante problematizada entre linguistas, pode ser entendida como a menor unidade não-presa dotada de sentido e que pode sofrer alterações morfológicas, como a flexão. O lexema, por sua vez, é uma unidade utilizada para fins de análise linguística e seu uso focaliza a base semântica da palavra, subfocalizando sua realização concreta (com flexões, por exemplo). Já a Unidade Lexical representa uma relação entre forma linguística e sentido tal que uma mesma forma pode evocar significados diferentes, a depender do esquema conceptual em que está inserida. A UL pode referir-se a uma única palavra gráfica ou a um grupo de palavras que atuam de forma conjunta. Assim, a partir do lexema *abrir*, é possível instanciar

as palavras *abriu*, *aberto*, *abrira*, *abríamos*. Quando o lexema é emparelhado com um frame é referido como uma Unidade Lexical evocadora desse frame. O lexema *abrir* pode ser uma UL de `Closure` (`Fechamento`), em *ele abriu a tampa da panela*; ou uma UL de `Body_movement` (`Movimento_corporal`), em *ele abriu bem os olhos*.

A descrição lexicográfica proposta busca ligar a significação dos lexemas em termos dos frames por eles evocados. Trata-se de descrever as Unidades Lexicais relevantes que, a depender do frame ao qual estejam emparelhadas, apresentam distintas possibilidades combinatórias, no âmbito semântico e sintático. Segundo o aporte da FrameNet, a descrição desses padrões de combinação, chamados de Valências, visa a identificar os Elementos de Frame (papéis semânticos específicos), e as realizações sintáticas desses Elementos, que figuram como a Função Gramatical (FG) desempenhada e o Tipo de Sintagma (TS) por eles representados. No Projeto FrameNet, a anotação mais praticada (e a adotada neste trabalho) é chamada *anotação lexicográfica*. Neste caso, o foco é elencar todas as valências de uma UL. Ou seja, a meta é "registrar todas as possibilidades semânticas e sintáticas (valências) de cada lexema em cada um de seus sentidos" (RUPPENHOFER *et al*, 2006, p. 20). Para isso, são extraídas sentenças de vastos *corpora* que abranjam textos de gêneros diversos.

3.2 Procedimentos para a Anotação Lexicográfica no Projeto FrameNet

Uma vez definido o frame e seus Elementos constitutivos, são definidas as ULs que evocam tal frame. O procedimento de anotação requer que se registrem as instâncias da palavra-alvo (a UL) em camadas específicas, onde serão explicitadas as valências dessa UL. Postulam-se no mínimo quatro camadas de anotação, que são as mais proeminentes: Camada da Palavra-Alvo; Camada dos Elementos de Frame (EF); Camada da Função Gramatical (FG) e Camada do Tipo Sintagmático (TS). A **Camada da Palavra-Alvo** contém a sentença na qual figura a Unidade Lexical e é onde a UL é destacada. Apenas uma palavra-alvo é estabelecida por sentença. As três próximas camadas relacionam-se intimamente com a valência da UL. A contraparte semântica da valência de uma UL são os Elementos de Frame que configuram suas diversas instâncias. A contraparte sintática se refere à classificação dos EFs quanto à Função Gramatical

exercida e o Tipo Sintagmático correspondente. Nestas camadas não há marcações para a UL, apenas para os Elementos de Frame.

É na **Camada dos Elementos de Frame** que os diversos constituintes da sentença são etiquetados. Os EFs centrais e não-centrais são identificados. A etiquetagem pode ser feita atribuindo-se (arbitrariamente) a cada elemento uma cor. Assim, todos os constituintes que instanciam determinado EF são marcados com a mesma cor. Caso um Elemento Central não se instancie lexicalmente na sentença, sua presença é inferível. Dessa forma, esse EF é anunciado com uma etiqueta de Instanciação Nula, que poderá ser Definida, Indefinida ou Construcional.

O frame `Fechamento` tem como Elementos Centrais o AGENTE, o PRENDEDOR e o OBJETO_TIPO_CONTÊINER. Uma vez que esses EFs são participantes centrais do frame `Fechamento`, todos devem ser depreendidos em todas as instâncias desse frame. Por exemplo, tem-se a seguinte sentença ilustrativa do frame `Fechamento` evocado pela UL *abrir_((tampa))*: "Traga a panela. –Agora abra a tampa". Assim fica a etiquetagem na Camada EF:

1- Traga a panela. –[Agora $_{\text{TEMPO}}$] ABRA [a tampa $_{\text{PRENDEDOR}}$] [AGENTE $_{\text{INC}}$] [OBJETO_TIPO_CONTÊINER $_{\text{IND}}$]

No referido exemplo, a omissão do AGENTE é licenciada pela Construção Imperativa e o OBJETO_TIPO_CONTÊINER, embora não esteja explícito na predicação de *abrir*, é recuperado anaforicamente pelo contexto. Observando a sentença, percebemos que o objeto do qual a tampa é deslocada é **a panela**. No entanto, o constituinte "a panela" pertence à predicação de *trazer* e não de *abrir* e, por isso, não é marcado.

Pode haver ainda, na Camada Elementos do Frame, dois fenômenos. Um deles relaciona-se à palavra-alvo e chama-se *incorporação*. Ocorre quando a Unidade Lexical e algum Elemento de Frame se fundem morfologicamente. Em quatro ULs descritas neste trabalho há esse fenômeno. Nas ULs *desabotoar*, *desarrolhar*, *desatarraxar* e *destampar*, os verbos (ULs) têm em si o EF PRENDEDOR *botão*, *rolha*, *tarraxa* e *tampa*, respectivamente. Outro fenômeno é o de *conflação*, que é o fato de haver, em um único constituinte, mais de um Elemento de Frame. O caso de *conflação* pode ser observado no exemplo "Vamos, Luca. Desabotoe sua camisa", do frame `Fechamento`. Neste caso, o

sintagma "sua camisa" contempla o EF central OBJETO_TIPO_CONTÊINER. No entanto, dada a presença do Possessivo (sua) dentro desse sintagma, é possível depreender o EF não-central POSSUIDOR (você).

Na **Camada da Função Gramatical** será marcada a Função Gramatical desempenhada pelos Elementos do Frame. De acordo com a classe gramatical da Unidade Lexical são determinadas possíveis Funções Gramaticais. No caso de uma UL verbal, a FrameNet postula três tipos diferentes de Função Gramatical (FG). Essas funções refletem, basicamente, a grade argumental do predicado. A anotação da FrameNet é localizada no escopo da predicação e as funções previstas são: Argumento Externo (Ext), Objeto (Obj) e Dependente (Dep).

A **Camada dos Tipos Sintagmáticos** é usada para identificar os Sintagmas relacionados à UL como Elementos de Frame. Isso quer dizer que, diante de uma sentença que instancie uma UL, apenas os constituintes que pertençam ao escopo da predicação são anotados. Também de acordo com a classe gramatical da UL, diferentes serão os possíveis Tipos Sintagmáticos. Os Sintagmas mais proeminentes são: Sintagmas Nominais (SN), Sintagmas Adjetivais (SA), Sintagmas Preposicionais (SP), Sintagmas Verbais (SV), Sintagmas Adverbiais (SAdv) e Orações (Or). É importante frisar que a FrameNet marca todo o segmento do Sintagma e não apenas seu Núcleo. Por exemplo, tem-se a sentença:

2-[Tenoch $_{\text{AGENTE/EXT/SN}}$] ABRIA [a tampa $_{\text{PRENDEDOR/OBJ/SN}}$] [do vaso $_{\text{OBJETO\_TIPO\_CONTÊINER/DEP/SP}}$] [com o pé $_{\text{MANIPULADOR/DEP/SP}}$] [na casa de Júlio $_{\text{LUGAR/DEP/SP}}$]

No caso acima, apesar de o EF OBJETO_TIPO_CONTÊINER ser *"vaso"*, é marcado como Sintagma Preposicional, visto que seu constituinte é *"do vaso"*. Também o EF MANIPULADOR da ação de abertura é "*pé*"; no entanto, o Sintagma inteiro "*com o pé*" é marcado e identificado como um Sintagma Preposicional (SP).

3.2.1 O sumariamento das Unidades Lexicais

Uma vez concluído o trabalho de anotação dos Elementos de Frames com suas respectivas Funções Gramaticais e identificação de seus Tipos Sintagmáticos, a FrameNet lança mão de um recurso, o *FrameNet Desktop*, que automaticamente faz o sumariamento da UL

em questão sob dois vieses: (i) realização individual de EFs e (ii) realização e interação dos EFs entre si.

No primeiro prisma, a FrameNet sumariza **todas** as instanciações de todos os Elementos de Frame que ocorreram nas sentenças extraídas dos *corpora*. Neste sumariamento, são contabilizados Elementos Nucleares e Não-nucleares. As instanciações nulas também são contabilizadas. Assim, é possível visualizar como cada EF em particular se realiza sintaticamente quando ocorre com uma determinada Unidade Lexical.

O segundo tipo sumariza todos os padrões de uso da UL. Os padrões de uso são os padrões de valências, sendo a combinação das três camadas mais relevantes (EF, FG e TS). Neste padrão não há interesse em apresentar a ordem linear de ocorrência, e sim em saber qual a integração de Elementos de Frame entre si e com suas realizações sintáticas (FG e TS).

3.3 Escolha das Unidades Lexicais

O interesse em pesquisar eventos de **separação física** se fundamenta por buscar integrar a presente investigação a um interesse em pesquisa lexicográfica de expressividade mundial. Eventos de *cortar* e *quebrar* foram tema de uma edição especial da revista *Cognitive Linguistics* no ano de 2007, que reuniu uma coletânea de quinze artigos sobre pesquisas feitas acerca das diferentes categorizações que tais eventos recebiam ao longo de diversas línguas. A escolha do grupo de ULs tanto do frame `Fechamento`, quanto do frame `Movimento_corporal` foi baseada em um trabalho investigativo realizado no Instituto Max Planck,[5] sobre a representação linguística dos eventos de descontinuidade de uma unidade física relacionados; a saber, 'cortar' e 'quebrar' (MAJID *et al*, 2007).

Foi formulado um esquema de agrupamento desses eventos dentro de um grande domínio de **separação física**. Este domínio se divide em dois tipos de separação física com base no critério da destruição material: (I) separação física **com perda material** da entidade; e (II) separação física **sem perda material** da entidade. Este

[5] http://www.mpi.nl/

segundo tipo compreende as ações de *abrir* e se caracteriza por promover o afastamento de partes de um todo. São casos de separação reversível. O primeiro tipo de separação física ainda é subdividido em dois outros grupos, levando-se em conta o critério da preditividade das partes resultantes da separação: (Ia) tipos de separação em que as **partes resultantes são preditíveis**, exemplificado por eventos de *cortar*; e (Ib) tipos de separação em que **as partes resultantes são impreditíveis**, exemplificado por eventos de *quebrar*.

4 A constituição e o tratamento dos *corpora*

Os *corpora* utilizados para esta investigação constituem o *Corpus* do Projeto FrameNet Brasil. Este, ainda em fase de expansão, é formado, atualmente, por cinco *corpora* eletrônicos que contemplam a variante do português do Brasil. Composto de aproximadamente 72 milhões de *tokens*, o *Corpus* FrameNet Brasil abrange diversos gêneros textuais. A seguir estão listados os *corpora* com suas respectivas descrições, contidas no portal FrameNet Brasil:[6]

(i) ANCIB: *corpus* criado a partir de mensagens enviadas para a lista homônima da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (até Novembro de 2003) e para a lista abarreto-l, após essa data;

(ii) ECI-EBR: *corpus* criado pela ECI (European *Corpus* Initiative), baseado no Borba-Ramsey. É uma seleção de excertos de obras brasileiras, contendo pelo menos discurso literário, didático e oral cuidado (discursos políticos);

(iii) LF (Legendas de Filmes): *corpus* criado pelo Projeto FrameNet Brasil, sediado na Universidade Federal de Juiz de Fora, contém legendas de filmes em Português do Brasil cedidas pelo portal OpenSubtitles.org;

(iv) NILC/São Carlos: contém textos brasileiros do registro jornalístico (do qual se originou o CETENFolha), didático, epistolar e redações de alunos;

[6] Acesso ao *Corpus* FrameNet Brasil na internet: <http://www.framenetbr.ufjf.br/index.php?option=com_content&view=article&id=9&Itemid=13&lang=pt>

(v) NURC-RJ: *corpus* constituído por entrevistas gravadas nas décadas de 1970 e 1990, num total de 350 horas, com informantes de nível superior completo, nascidos no Rio de Janeiro e filhos de pais preferencialmente cariocas.

Para cada UL descrita, é feita a busca em cada um desses cinco *corpora*. Três deles são disponíveis ao acesso público em rede, na página da Linguateca,[7] a saber: ANCIB, ECI-EBR e Nilc/São Carlos. Na referida página é realizada a busca lematizada das palavras-chave. Dentre os outros dois *corpora*, o LF (Legenda de Filmes), criado pelo Projeto FrameNet Brasil e o NURC-RJ, disponibilizado pelo Projeto Norma Linguística Urbana Culta- RJ, estão alojados na página do Sketch Engine,[8] onde é feita a busca, ainda não lematizada, das palavras-chave.

4.1 Busca eletrônica

A busca nos *corpora* diferenciou-se quanto ao caráter mono ou polilexêmico das expressões linguísticas constitutivas das ULs selecionadas. Houve um esquema de busca para as expressões monolexêmicas: *desabotoar*, *desarrolhar*, *desatarraxar* e *destampar*, todas pertencentes ao frame `Fechamento`; e outro esquema de busca para as expressões polilexêmicas *abrir_((tampa))*, *tirar_((tampa))*, *levantar_((tampa))*, do frame `Fechamento`, e *abrir_((boca))*, *abrir_((mão))* e *abrir_((olho))*, do frame `Movimento_corporal`. Outra diferenciação nas buscas deveu-se à questão da lematização dos *corpora*. Foi, portanto, sensivelmente diferente o modo de proceder à busca nos *corpora* ANCIR, ECI-EBR e Nilc/São Carlos (lematizados[9]) da busca em LF e NURC-RJ (ainda não lematizados).

[7] Linguateca- página inicial de visualização dos *corpora* disponíveis < http://www.linguateca.pt/ACDC/>

[8] Sketch Engine- página < http://www.sketchengine.co.uk/>

[9] Um *corpus* lematizado é aquele que passou por um tratamento sintático e no qual é possivel obter as flexões verbais a partir do lema do verbo. Assim, uma busca por "**abrir**" retorna resultados como **abr**iu, **abr**o, **abr**indo, **ab**erto, etc.

4.1.1 Expressões monolexêmicas

Em *corpora* lematizados (ANCIB, ECI-EBR e NILC/São Carlos), a busca por expressões monolexêmicas foi feita a partir da fórmula: [lema="*verbo na forma infinitiva*"], inserida na caixa de busca. O lema é a forma inflexionada da palavra. Assim, a partir da fórmula [lema="desabotoar"], é possível elencar todas as diferentes flexões desse verbo disponíveis no *Corpus* em questão. O resultado de busca visado é o que apresente a concordância da palavra-chave, ou seja, as diversas formas da palavra relacionadas a diferentes contextos linguísticos.

A diferença de busca em um *corpus* não lematizado é que, neste caso, a busca é feita por sequência de caracteres. Dessa forma, se se deseja pesquisar "desabotoar" em suas diversas flexões, a busca é feita por meio da fórmula "(?i)desaboto.*", onde <<(?i)>> atua como código para indistinção entre letras maiúsculas e minúsculas, e <<.*>> atua como código para que sejam incluídos quaisquer caracteres após a sequência *desaboto*.

4.1.2 Expressões polilexêmicas

Para se conseguir abarcar os contextos de uso de expressões multilexêmicas, a busca deve ser feita de duas formas. Por exemplo, para a UL *Abrir_((boca))*, há a possibilidade de diferentes ordens lineares entre os constituintes *abrir* e *boca*. Por exemplo, precisaríamos abarcar casos como "*esse aí só* ***abre*** *a* ***boca*** *para mastigar mesmo*" (o verbo *abrir* aparece antes de *boca*), bem como instâncias de "*-O que tem na* ***boca****?* ***Abra****. Deixe-me ver*" (*boca* aparece antes do verbo *abrir*). Além disso, seria necessário prever uma margem máxima de material linguístico entre *abrir* e *boca* (e vice versa), e, por isso, foi definido que essa margem seria de vinte palavras. Para dar conta de tais especificidades, o modo de buscar expressões multilexêmicas em nossos *corpora* seguiu a seguinte fórmula (usamos *Abrir_((boca))* como exemplo): <[lema="abrir"] [] {0,20} "boca.*"> e <"boca.*" [] {0,20} [lema="abrir"] >.

Como se percebe, este tipo de busca é aplicado a *corpora* lematizados. E para os *corpora* ainda não lematizados a fórmula[10] de busca é alterada para: <"(?i)abr.*" [] {0,20} "(?i)boca.*"> e <"(?i)boca.*" [] {0,20}"(?i)abr.*">.

Assim, para cada UL que envolve uma pesquisa por expressões polilexêmicas são feitas duas buscas e os resultados são agrupados. A seguir, listamos as fórmulas de busca utilizadas nesta pesquisa:

		Unidade Lexical	***Corpora* lematizados**	***Corpora* não lematizados**
A	Mono	*Desabotoar* Desarrolhar Desatarraxar *Destampar*	[lema="desabotoar"] [lema="desarrolhar"] [lema="desatarraxar"] [lema="destampar"]	"(?i)desaboto.*" "(?i)desarrolh.*" "(?i)desatarrax.*" "(?i)destamp.*"
	Poli	*Abrir_((tampa))*	[lema="abrir"] []{0,20} "tamp.*" "tamp.*" []{0,20}[lema="abrir"]	"(?i)abr.*" []{0,20}"tamp.*" "tamp.*"[]{0,20} "(?i)abr.*"
		Levantar_((tampa))	[lema="levantar"][]{0,20}"tamp.*" "tamp.*"[]{0,20}[lema="levantar"]	"(?i)levant.*" []{0,20}"tamp.*" "tamp.*"[]{0,20}(?i)levant.*"
		Tirar_((tampa))	[lema="tirar"] []{0,20} "tamp.*" "tamp.*" []{0,20}[lema="tirar"]	"(?i)tir.*"[]{0,20}"tamp.*" "tamp.*" []{0,20} "(?i)tir.*"
B	Poli	*Abrir_((boca))*	[lema="abrir"] []{0,20} "boca.*" "boca.*" []{0,20}[lema="abrir"]	"(?i)abr.*" []{0,20}"boca.*" "boca.*"[]{0,20} "(?i)abr.*"
		Abrir_((mão))	[lema="abrir"] []{0,20} "mão.*" "mão.*"[]{0,20} (?i)abr.*"	"(?i)abr.*"[]{0,20}"mão.*" "mão.*"[]{0,20}[lema="abrir"]
		Abrir_((olho))	[lema="abrir"] []{0,20}"olh.*" "olh.*" []{0,20}[lema="abrir"]	"(?i)abr.*"[]{0,20}"olh.*" "olh.*" []{0,20} "(?i)abr.*"

A= `Fechamento`

B= `Movimento_corporal`

[10] Por restrições metodológicas, os caracteres de busca (visando resultados para o verbo ABRIR) tiveram de ser "abr.*". Isso não nos possibilitou obter resultados como "aberto/a", o que, dada a possibilidade de construções passivas, significa uma perda em termos de variabilidade de ocorrências. No entanto, para abarcar resultados de ABRIR no particípio, em *corpora* não lematizados, a busca deveria ser "ab.*", o que traria uma gama de resultados indesejáveis, como "**ab**ençoar", "**ab**orrecimento", "**ab**rigo", etc.

4.2 Tratamento dos *subcorpora*

É importante frisar que a busca eletrônica em *corpora* é feita pelo lexema e não pela Unidade Lexical, o que torna imprescindível proceder à limpeza dos *corpora*, disponibilizando para a anotação lexicográfica apenas as ocorrências que correspondam aos frames pesquisados.

Para que a limpeza seja feita de forma criteriosa, é utilizado um programa computacional, o ***Tinn-R***, a partir do qual é possível visualizar, de forma estatística, as classificações dos diversos usos do lexema nas sentenças extraídas dos *corpora*. A funcionalidade do Programa ***Tinn-R*** para nossa análise é proporcionar o refinamento sistemático dos *corpora*. No entanto, essa funcionalidade se mostra necessária para casos de *corpora* com um número grande de ocorrências, visto que a tarefa de agrupamento de usos, a olho nu, seria exaustiva e passível de erros. Diante de um número de dez ULs (que implica uma busca por dez expressões linguísticas) a serem investigadas em cinco *corpora*, há a possibilidade de serem geradas cinquenta planilhas do Excel.

Em nossa busca, entretanto, houve casos de buscas que não retornaram ocorrência alguma. Outros casos houve em que o número de ocorrências não ultrapassava quatro ou cinco sentenças. Dessa forma, foi estabelecido que, apenas as buscas que retornassem acima de **cinco** ocorrências (quer válidas ou não) seriam submetidas ao Programa ***Tinn-R***. A tabela abaixo mostra que foram geradas vinte planilhas do Excel para todo o trabalho:

Fechamento		
Unidade Lexical	**Nº Planilhas geradas**	***Corpora***
Desabotoar	2	LF, NILC/ São Carlos
Desarrolhar	0	
Desatarraxar	0	
Destampar	0	
Tirar_((tampa))	1	LF
Abrir_((tampa))	3	NILC/São Carlos, ECI-EBR, LF
Levantar_((tampa))	1	LF
Movimento_corporal		
Unidade Lexical	**Nº Planilhas geradas**	***Corpora***
Abrir_((olho))	4	NILC/São Carlos, ECI-EBR, NURC-RJ, LF
Abrir_((boca))	4	NILC/São Carlos, ECI-EBR, NURC-RJ, LF
Abrir_((mão))	5	Todos
Total	20	

Planilhas geradas no Excel para depreensão dos *subcorpora*

Embora fossem raros, houve casos em que a busca retornou um número muito grande de ocorrências, o que dificultaria a análise de filtragem das sentenças. Assim, dado o escopo da pesquisa, foi estabelecido um número máximo de 400 ocorrências por *corpus*. Quando esse número fosse ultrapassado, o procedimento metodológico era submeter as ocorrências totais do *corpus* a um processo de amostragem. Para este fim, foi utilizado o SPSS (**S**tatistical **P**ackage for the **S**ocial **S**ciences),[11] que é um *software* estatístico utilizado para realizar a amostragem de dados.

Os resultados das buscas para cada lexema em cada *corpus* são copiados e colados em planilhas do Microsoft Office Excel. Cada uma das vinte planilhas é o *subcorpus* de cada lexema que será submetido a um refinamento, onde serão separados os diversos usos.

O procedimento para utilizar o recurso do Programa ***Tinn-R*** é atribuir a cada sentença (na planilha do Excel) uma determinada Classificação, de acordo com o uso do lexema. Para que o programa ***Tinn-R*** possa "ler" a Classificação atribuída à sentença são utilizados números correspondentes à classificação que se queira dar. A Classificação para as sentenças é feita na coluna B, a coluna da "Classificação". No quadro abaixo, é apresentada a Classificação referente a cada número, que corresponde ao *script* executado pelo Programa ***Tinn-R***:

Número/Código	**Classificação**
1	Sentido físico
2	Sentido figurativo
3	Adjetivo
4	Substantivo
5	Contexto ambíguo ou insufiente
6	Outros

Scrip para o ***Tinn-R*** das Classificações atribuídas aos Lexemas pesquisados

[11] Pacote Estatístico para as Ciências Sociais. Maiores informações podem ser obtidas acessando a página Wikipédia: http://pt.wikipedia.org/wiki/SPSS.

1- **Sentido físico**: classificação usada para os casos em que o lexema for empregado como Unidade Lexical que evoque o frame pesquisado.

2- **Sentido figurativo**: classificação usada para os casos em que o lexema for empregado com uso figurativo (metafórico, metonímico, etc).

3- **Adjetivo**: classificação usada para os casos em que o lexema, apesar de empregado como uma UL do frame pesquisado, estiver sob a forma adjetival. Por exemplo, no caso de "blusa desabotoada".

4- **Substantivo**: classificação usada para os casos em que o lexema, apesar de empregado como uma UL do frame pesquisado, estiver sob a forma substantival. Por exemplo, num caso como "o desabotoamento".

5- **Contexto ambíguo**: classificação usada para os casos em que não for possível depreender o sentido empregado pelo lexema, dada a escassez ou ambiguidade do material circunvizinho.

6- **Outros**: classificação usada para casos em que o lexema estiver empregado em cenas diferentes das de separação física (e que não seja de maneira figurada). Esta classificação tanto serve para outros frames evocados como também para casos em que houver erros de digitação.

Uma vez feita a classificação de todas as sentenças, que é a primeira parte analítica do processo lexicográfico que empreendemos, o arquivo do Excel é submetido ao processo estatístico do Programa ***Tinn-R***. A partir do *script* inserido (com as Classificações de 1 a 6), o programa disponibiliza, em termos absolutos e percentuais, os diferentes usos do Lexema.

Após a execução do *script*, o programa gera um arquivo no formato txt com as frases referentes apenas ao tipo de classificação solicitado. No nosso caso, é retornado apenas o conjunto de frases classificadas com o sentido 1 – físico e em forma verbal. Os arquivos em txt gerados pelo Programa ***Tinn-R*** para cada *corpus* são agrupados em um único documento de texto (no nosso caso, o Microsoft Word) e, assim, é formado o banco de dados com as sentenças que contemplam

os empregos da UL em gêneros textuais diversos. A partir daí segue-se o todo o procedimento de anotação das ULs, que constitui a principal parte analítica deste trabalho: a descrição lexicográfica.

5 Alguns apontamentos sobre a análise de uls do frame `Fechamento`

Inserida no evento maior e mais genérico de separação física, a cena de abertura (e fechamento) de partes componentes de uma unidade física é bem exemplificada no frame `Fechamento`. Pela forma como é descrito, esse esquema conceptual se refere à ação de *abrir* ou *fechar* algum objeto. Sob esse viés, caracteriza-se como um frame agentivo.

A seguir, apresentamos de forma resumida a definição do frame `Fechamento` aos moldes de como é feita a definição de frames pela FrameNet americana. No entanto, dado o escopo deste trabalho, elencamos apenas os Elementos Centrais:

Frame Fechamento
Definição:
Um Agente manipula um Prendedor para abrir ou fechar um Objeto_tipo_contêiner (por exemplo: casaco, jarra). Algumas vezes uma ou um Portal_do_contêiner podem ser expressos. Uma vez que o Manipulador é sintaticamente omissível, vários verbos neste frame incorporam o Prendedor.

Elementos de Frame Nucleares:

Agente	O Agente abre/fecha o Objeto_tipo_contêiner.
Portal_do_Contêiner	Ela DESTAMPOU a garrafa Este EF identifica a zona ativa do contêiner que pode ser expressado como o Objeto Direto do Verbo alvo. D. Odete DESABOTOAVA a braguilha.
Objeto_tipo_contêiner	Este EF identifica o item que é fechado pelo Agente com o Prendedor.
Prendedor	Este EF identifica o Prendedor que o Agente manipula. Você tem que TIRAR a tampa da lente Ela DESTAMPOU a garrafa.

São apresentados alguns resultados pontuais em relação ao procedimento de descrição lexicográfica empreendido. A comparação entre os números de ocorrências de lexemas encontrados na busca eletrônica e as ULs depreendidas a partir desses lexemas é apresentada a seguir:

Frame `Fechamento`			
	Ocorrências totais (Lexemas)	Ocorrências válidas (ULs)	Percentual de ULs
ULs monolexêmicas			
Desabotoar	68	46	67%
Desarrolhar	03	01	33%
Desatarraxar	04	03	75%
Destampar	13	06	46%
Subtotal	88	56	64%
ULs polilexêmicas			
Abrir_((tampa))	69	31	45%
Levantar_((tampa))	36	16	44%
Tirar_((tampa))	58	20	34%
Subtotal	163	67	43%
Total em `Fechamento`	251	123	49%

Como se percebe, os resultados de busca por lexemas (no caso dos monolexêmicos) e por lexemas em interação (no caso dos polixelêmicos), não chegaram a setenta ocorrências em nenhum dos casos. Analisando sob outro viés, a quantidade de resultados válidos oscilou menos entre as buscas por monolexêmicos do que entre os polilexêmicos. De forma holística, para o frame Fechamento, o resultado da busca por lexemas foi de 251 ocorrências, e, desse total, pouco menos da metade foi validada como ULs referentes ao frame Fechamento. A outra metade, não validada, referia-se, na maioria dos casos, aos lexemas que figuravam como ULs vinculadas a outros frames, ou mesmo que foram empregados nas sentenças sob forma figurativa.

A seguir, são feitos comentários sobre as Unidades Lexicais individualmente, iniciando pelas ULs monolexêmicas. Em *Desabotoar*, por se tratar de uma UL em que há incorporação do EF PRENDEDOR, é previsto que este seja depreendido no próprio Verbo; como, de fato, ocorreu em 40 dos 46 casos. Nos seis casos em que o EF PRENDEDOR foi explicitado, cinco deles serviram para enfatizar o elemento **botão**, como em "*Ponha os cabelos sobre os olhos e* DESABOTOE [os botões OBJETO_TIPO_CONTÊINER]". Nesse caso, o alvo *Desabotoar* parece estender seu sentido para o ato de *abrir*, não se limitando ao ato de *desapertar botões*.

A UL *Desarrolhar* apresentou apenas uma sentença, com AGENTE não expresso linguisticamente e EF PRENDEDOR incorporado. O fato de este número não ser representativo nos impede de formular generalizações quanto ao padrão sintático. A ocorrência "*Depois,* DESARROLHAMOS um [Dom Perignon OBJETO_TIPO_CONTÊINER]" revela o uso de *Desarrolhar* exclusivamente para representar a abertura de garrafa. O fato de se tratar de uma bebida fina (champanhe) sugere que o uso desta UL se relacione a uma evocação de cena mais formal, onde a abertura do objeto figure como ação central e, portanto, mais provável de ser lexicalizada. Possivelmente esta UL não seja proeminente no português falado no Brasil, em que a UL *Abrir* (*a garrafa*/ *a tampa da garrafa*) parece ser mais usada.

A UL *Destampar* apresentou seis sentenças, dispostas em quatro padrões sintáticos. Quanto à realização lexical do EF PRENDEDOR, houve homogeneização nesta UL, uma vez que todos os casos foram de incorporação. Em relação ao EF OBJETO_TIPO_CONTÊINER, sua realização, na maior parte, como Objeto sintático foi considerada canônica. O AGENTE manifestou-se de forma mais heterogênea, embora tenha ocorrido mais como Instanciação Nula. Nos casos de AGENTE como INI foi observado o contexto culinário, como em: "*Quando começar a ferver, marcar vinte minutos e deixar cozinhar o abacaxi, sem* DESTAMPAR [a panela OBJETO_TIPO_CONTÊINER]".

As ULs polilexêmicas, todas empregadas como verbos transitivos, tendem a apresentar o EF PRENDEDOR (*tampa*, em todos os casos) como Objeto direto. Já o EF OBJETO_TIPO_CONTÊINER é previsto para figurar sintaticamente como um Dependente e sintagmaticamente como um Sintagma Preposicionado. O AGENTE é, por sua vez, esperado de ocorrer como Argumento Externo do verbo. No entanto, essa configuração foi observada em somente cinco casos para a UL *Tirar_((tampa))*, dois casos para a UL *Levantar_((tampa))* e em apenas um caso para a UL *Abrir_((tampa))*.

A UL *Abrir_((tampa))* apresentou 31 sentenças, dispostas em 11 padrões sintáticos. O EF PRENDEDOR ocorreu na maior parte dos padrões como Objeto direto de *abrir*, totalizando 28 das 31 sentenças. A instanciação mais recorrente do EF OBJETO_TIPO_CONTÊINER foi como IND, distribuída em metade dos padrões. O EF AGENTE ocorreu como instanciação nula em 26 sentenças. Dentre essas, em 16 casos (encontrados em dois padrões sintáticos) foi instanciado como IND, ou seja, inferido pelo contexto e depreendido pela flexão verbal.

A UL *Levantar_((tampa))* apresentou 20 sentenças, dispostas em seis padrões sintáticos. O EF PRENDEDOR teve uma representação homogênea, pois consta em todas as instâncias como Objeto Direto do verbo *levantar*. O EF AGENTE teve realizações variadas. Embora tenha ocorrido mais frequentemente como Instanciação Nula do que lexicalizado sintaticamente, há uma oscilação em casos de IND e INC. Assim, cada realização do EF AGENTE distribuiu-se em dois padrões. Casos de IND remetem a situações de inferência contextual, já os casos de INC se referem aos casos de modo verbal imperativo.

UL *Tirar_((tampa))* apresentou 20 sentenças, dispostas em sete padrões sintáticos. A variação do EF AGENTE é equilibrada entre realização explícita e nula. Assim como *Levantar_((tampa))*, a UL *Tirar_((tampa))* apresentou o EF PRENDEDOR exclusivamente como Objeto Direto do alvo *tirar*. O EF OBJETO_TIPO_CONTÊINER é mais recorrente como IND. No entanto, chamamos a atenção para o único caso em relação a todas as ULs polilexêmicas do frame `Fechamento` em que este EF ocorreu como Sintagma Nominal (por meio de pronome pessoal), percebido no trecho: "*Tomamos o tubo,* TIRAMO-[lhe $_{\text{OBJETO\_TIPO\_CONTÊINER}}$] [a tampa $_{\text{PRENDEDOR}}$]".

6 Alguns apontamentos sobre a análise de uls do frame `Movimento_corporal`

Por ser um esquema conceptual bastante genérico (basta imaginar a gama de movimentações possíveis de serem concretizadas com as várias partes do corpo), o frame `Movimento_corporal` nos interessou por apresentar, dentre suas várias especificidades, o evento de **afastamento de partes contíguas**. Esse afastamento é caracterizado também como um ato de **abertura**. Pretendemos descrever cenas de abertura de olhos, da boca e da mão e verificar possíveis traços de semelhança entre esses movimentos e o evento de *abrir* e *fechar* objetos.

A seguir, há a descrição do frame `Movimento_corporal` com foco nos EFs centrais:

Frame Movimento_corporal
Definição:
Este frame contém palavras para movimentos ou ações que um Agente desempenha usando algumas partes de seu corpo.
Elementos de Frame:
Nucleares
Agente [Agt] O Agente usa alguma parte do seu corpo para desempenhar uma ação.
Ezequiel ABRIU a boca.

É a entidade movimentada.Geralmente ocorre como Objeto Direto.
ela não ABRIU !

Em relação ao procedimento de descrição lexicográfica das três ULs polilexêmicas do frame `Movimento_corporal`, apresentamos também alguns resultados pontuais. A comparação entre os números de ocorrências de lexemas encontrados na busca eletrônica e as ULs depreendidas a partir desses lexemas é apresentada a seguir:

Frame Movimento_corporal			
Ocorrências totais (Lexemas)	Ocorrências válidas (ULs)	Percentual de ULs	
ULs polilexêmicas			
Abrir_((boca))	497	104	21%
Abrir_((mão))	650	08	1%
Abrir_((olho))	612	190	31%
Total em `Movimento_corporal`	1756	302	17%

O número de lexemas retornados foi mais expressivo neste frame, se comparado ao `Fechamento`. No entanto, o aproveitamento dos dados foi relativamente pequeno (17%) e significativamente menor que em `Fechamento` (que apresentou 49% de aproveitamento).

Os casos mais numerosos foram de realização dos lexemas ou como forma **Figurativa** ou como **Outros** (quando, na maioria das vezes, a presença de *abrir* + *boca/mão/olho* na ocorrência se relaciona a cenas e situações diversas). Em relação à classificação como **Outros**, houve casos como: "*A voz ia ficando lírica, dava de se entregar pelos*

olhos *ao mulato Flodoaldo. Acostumado aqueles sintomas, o barraqueiro começava a* ***abrir*** *o jogo*", retirado no *corpus* ECI-EBR para a busca <[lema="abrir"] [] {0,20}"olho.*">, em que *abrir* e *olho* sequer ocorrem na mesma sentença, referindo-se a situações distintas.

O grande número de ocorrências no sentido **Figurativo** deve-se ao fato de as expressões **abrir a boca**, **abrir mão** e **abrir o olho** serem bastante produtivas enquanto significação de **falar**; **renunciar**; e **prestar atenção** respectivamente. Como ilustração, citamos o caso retornado da busca <[lema="abrir"] [] {0,20}"boca.*"> no *corpus* Legenda de Filme: "*Ninguém tem que provar o contrário. O dever de provar é da procuradoria. -O réu nem mesmo tem que* ***abrir*** *a* ***boca****. Está na constituição. -Claro, sei disso*", em que *abrir a boca* é **falar**; e a ocorrência pela busca <[lema="abrir"] [] {0,20}"mão.*"> no *corpus* Ancib: "*Autores* ***abrindo mão*** *de seus direitos devem ler cuidadosamente as letras pequenas dos novos contratos de acesso livre*", quando *abrir mão* é empregado com sentido de **renunciar**.

Em relação à descrição lexicográfica, consideramos os EFs AGENTE e PARTE_DO_CORPO em termos de suas realizações lexicalmente expressas ou como Instanciações Nulas. Dada a configuração das ULs verbais, que têm o verbo *abrir* em interação com alguma parte corporal (*boca*, *mão* ou *olho*), a tendência do EF PARTE_DO_CORPO é realizar-se como um sintagma nominal, Objeto Direto de *abrir*. De fato, esta foi a configuração em aproximadamente 87% de todas as ocorrências das três ULs. Quando não lexicalizado, este EF apresenta-se como IND, uma vez que a parte corporal é prevista na própria estrutura da UL e, portanto, caso não esteja no entorno sintático da palavra-alvo, estará presente no contexto.

A UL *Abrir_((boca))* apresentou 104 sentenças, distribuídas em 07 padrões sintáticos. A realização do EF AGENTE como Instanciação Nula representa 79% dos casos. Nesse contexto, a mais frequente realização é como INC, a qual se refere aos casos de modo verbal imperativo. Configurando ordem ou pedido, as situações de INC se relacionam ao ato de comer (em 25 das 66 sentenças), como em: " – ABRA [a boca $_{\text{PARTE\_DO\_CORPO}}$]. *Eu te dou de comer*"; e ao ato de tomar um remédio ou estar em consulta médica (em 15 das 66 sentenças), como em: "*-Deixe o doutor examiná-la. -Pare,* ABRA [a boca $_{\text{PARTE\_DO\_CORPO}}$]". Os outros casos foram variados ou considerados imprecisos.

Na maioria das vezes em que o EF PARTE_DO_CORPO é anotado como IND trata-se de casos de reiteração do evento. Isso implica que

houve menção anterior e a repetição parece funcionar como ênfase, como observamos em: "*-Abra a boca. Bem! -Vai ter que* ABRIR", em que há, nitidamente, a elipse de **boca**.

A UL *Abrir_((mão))* apresentou oito sentenças, distribuídas em quatro padrões sintáticos. A abertura das mãos, como ilustram as oito sentenças, não tem um evento específico mais recorrente. Refere-se, em alguns casos, ao movimento mecânico de *abrir* mãos, ou ao movimento intencional de *abrir* a mão com o objetivo de soltar ou visualizar algo escondido. A realização sintática do EF PARTE_DO_CORPO é mais homogênea e ocorre em sete dos oito casos.

A UL *Abrir_((olho))* apresentou 190 sentenças, distribuídas em oito padrões sintáticos. Esta foi a mais numerosa UL do frame `Movimento_corporal`. Apresenta, como mais expressivas, as realizações do EF PARTE_DO_CORPO como Sintagma Nominal e Objeto Direto e as realizações do EF AGENTE como Instanciação Nula.

As situações nas quais a UL *Abrir_((olho))* está inserida relacionaram-se, em 29 casos, ao evento de acordar ou despertar, como observamos em: "*ao acordar*, ABRINDO [os olhos $_{\text{PARTE\_DO\_CORPO}}$], *tudo voltaria a ser como antes*". Em 36 casos houve uma interação entre o ato de abrir os olhos e eventos que envolviam uma situação de surpresa, como é ilustrado em: "*– Tenho um presente para você. – Presente? – Feche os olhos... – E vire-se. – Pode* ABRIR [os olhos $_{\text{PARTE\_DO\_CORPO}}$]". Em 48 casos, a ocorrência desta UL remetia-se ao ato mecânico de abertura dos olhos, percebida em: "*Osmar continua em coma, mas está* ABRINDO [os olhos $_{\text{PARTE\_DO\_CORPO}}$] *espontaneamente por segundos*". Houve outros casos em que o ato de abrir os olhos refere-se a situações diversas ou imprecisas; e têm como unidade o fato de remeter-se à visão, como em: "*Querida Prudence por que* [você $_{\text{AGENTE}}$] *não* ABRE [seus olhos $_{\text{PARTE\_DO\_CORPO}}$]*? -Olhe em sua volta Olhe em sua volta*".

7 Considerações finais – com foco nas contribuições metodológicas do trabalho

Podemos, a partir deste trabalho, considerar alguns avanços em termos de refinamento metodológico de tratamento e apresentação dos dados. Inicialmente, o Projeto FrameNet Brasil considerava uma gama maior de *corpora*, pois abarcava também as variedades do português europeu. Optamos, ao dar prosseguimento à pesquisa com o presente trabalho, adotar unicamente os dados linguísticos da

variedade brasileira do português por considerarmos que muitas das especificidades entre as variedades brasileira e europeia se refletem em nuances diferences de padrões sintáticos, o que poderia ser um fator negativo no nosso objetivo de especificação e deliminitação dos usos linguísticos.

Outro ponto considerado um progresso metodológico foi a diversificação de gêneros textuais presentes nos *corpora* do Projeto FrameNet Brasil, com a implantação do *corpus* NURC-RJ e construção do *corpus* LF (Legendas de Filmes). Esses *corpora* propiciaram uma maior atestabilidade da variedade de língua falada, ou de escrita oralizada, e permitiram a obtenção de um número maior de instanciações das ULs evocadoras de separação física, que mostraram uma incidência pouco expressiva quando computados apenas os outros *corpora* constitutivos do Projeto.

Por se tratar de um processo analítico, o procedimento de etiquetação das ULs e Elementos de Frame é feito manualmente pela equipe de pesquisadores do Projeto FrameNet Brasil, assim como ocorre entre os envolvidos no Projeto Mãe. No entanto, no que tange à organização e ao sumariamento dos dados, esses são procedimentos ainda feitos manualmente por nós (ao passo que são disponíveis ferramentas computacionais para este fim no Projeto FrameNet americano): o que se traduz em um procedimento exaustivo e trabalhoso. Tais procedimentos manuais vêm contribuindo para o desenvolvimento de uma ferramenta computacional – a *FrameNet Brasil DeskTop* – que será capaz de sumariar os dados anotados e disponibilizá-los on line no site do nosso Projeto (www.framenetbr.ufjf.br); sendo essa ferramenta um grande avanço previsto.

Também ressaltamos que nossa proposta de não considerar como palavras-alvo os verbos *abrir*, *levantar* e *tirar* em si é ancorada no fato de assumirmos que a evocação da cena de separação física ocorre devido à presença não apenas dos verbos, mas quando estes estão em interação com EFs próprios da cena de Fechamento ou Abertura. Evidencia-se, então, o papel holístico das expressões, em detrimento do aspecto composicional de seus constituintes.

Uma das contribuições significativas destes trabalhos iniciais é atuarem como um laboratório para implementação, sofisticação e aprimoramento dos procedimentos envolvidos no Projeto FrameNet Brasil. Trabalhos como este, árduos e pedestres como possam parecer, resultam, na verdade, na construção de plataformas que em breve,

esperamos, possam contribuir de forma substantiva para o desenvolvimento de ferramentas computacionais de paráfrases, abreviamentos, traduções e busca semântica. Dessa forma, pretendemos que as análises feitas contribuam para a construção da contraparte para o português brasileiro da rede semântica FrameNet.

Referências

FILLMORE, C. J. Frame Semantics. In: *Linguistics in the morning calm. Selected papers from SICOL-1981*. Seoul, Korea: Hanshin Publishing Company, 1982.

FILLMORE, C.; JONHSON, C.; PETRUCK, M. Background to FrameNet. *International Journal of Lexicography*, v. 16, n. 3. Oxford University Press, 2003.

GAWRON, J. M. *Frame Semantics*. 2008. Disponível em: <http://www.hf.uib.no/forskerskole/new_frames_intro.pdf > Acesso em: 20 de janeiro de 2010.

MAJID, A. *et al*. The Semantic categories of cutting and breaking events: a crosslinguistic perspective. In: DABROWSKA, Ewa. (Ed.). *Cognitive Linguistics*. 2007. p. 133-152.

PETRUCK, M. *Frame Semantics*. University of California, Berkeley. 2008. Disponível em: <http://framenet.icsi.berkeley.edu/papers/miriamp.FS2.pdf> Acesso em: 20 de janeiro de 2010.

RUPPENHOFER, J. *et al*. *FrameNet II*: Extended theory and practice. Disponível em: <http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126> Acesso em: 20 de janeiro de 2010.

SALOMÃO, M. M. M. FrameNet Brasil: um trabalho em progresso. *Calidoscópio*, v. 7.2, 2009.

IV GRAMÁTICA E ESTRUTURA INFORMACIONAL

A unidade de Apêndice de Comentário – uma análise informacional a partir de dados do C-Oral-Brasil

Cássia Jacqueline Fernandes Oliveira[1]

RESUMO: Esse estudo consiste em uma análise baseada em corpora da unidade informacional de Apêndice de Comentário no PB. Essa pesquisa é sustentada pela Teoria da Língua em Ato (Cresti, 2000), segundo a qual um enunciado é definido como sendo a menor unidade passível de interpretabilidade pragmática. As fronteiras entre enunciados e suas unidades internas são delimitadas pela entonação. Uma das possíveis unidades internas é a unidade informacional de Apêndice de Comentário (APC), cuja função é a de estabelecer uma relação de integração com a unidade ilocucionária de COM (CMM ou COB). Pesquisas no italiano (Cresti, 2000; Tucci 2006) e no PB (Ulisses, 2008) mostraram que a unidade de APC apresenta um perfil nivelado e descendente e que possui médias de F0 e intensidade inferiores às médias da unidade ilocucionária a qual é integrada. Essa investigação foi conduzida em um subcorpus de PB constituído de 20 textos do C-ORAL-BRASIL, de aproximadamente 1500 palavras cada.

PALAVRAS-CHAVE: Estrutura Informacional; Atos de Fala; Apêndice de Comentário.

ABSTRACT: This study is a corpus-based analysis of the information unit Appendix of Comment in BP. This research is based on the *Informational Patterning Theory* (CRESTI, 2000), according to which an utterance is defined as the smallest unit allowing pragmatic interpretability. The boundaries between utterances and their internal units are delimited by intonation. One of the possible internal units is the information unit Appendix of Comment (APC), whose function is to establish a relation of integration with the illocutionary unit COM (CMM or COB). Research in Italian (Cresti, 2000, Tucci, 2006) and BP (Ulisses, 2008) showed that the APC has a flat and

[1] Doutora em Línguística Teórica e Descritiva. UFMG. E-mail: cassiajfo@gmail.com

falling movement and F0 and intensity values lower than the average of the illocutionary unit to which it is integrated. This investigation was performed on a BP subcorpus consisting of 20 texts of the C-ORAL-BRASIL, of approximately 1,500 words each.

KEYWORDS: informational patterning; speech acts; appendix of comment.

1 Introdução

Ao longo da história, vários pesquisadores se propuseram a estudar a linguagem, principalmente a escrita, deixando em segundo plano a fala. Ainda que semelhante em alguns aspectos, sabe-se que cada uma delas possui as suas especificidades e que analisar a fala pela lente da escrita é um equívoco.

Sem dúvida, uma questão muito discutida pela linguística atual é compreender como o falante organiza a informação na fala, isto é, como se organiza a sua estrutura informacional. A Teoria da Língua em Ato (CRESTI, 2000), que serviu de arcabouço teórico para esse estudo, foi desenvolvida para lidar com essas questões, inserindo o estudo da estrutura informacional dentro daquele dos atos de fala (AUSTIN, 1962).

2 A Teoria da Língua em Ato

A Teoria da Língua em Ato (Cresti, 2000) fundamenta-se em um estudo empírico da fala espontânea realizado pelo LABLITA (*Laboratorio Linguistico del Dipartimento di Italianistica dell'Università di Firenze*).

A fala espontânea, de acordo com essa perspectiva teórica, é considerada como toda a produção linguística sonora dialogada ou monologada em situação natural, realizada livremente, em contextos e situações comunicativas naturais, formais ou informais.

A imposição de um molde de segmentação do texto escrito sobre o discurso falado leva o pesquisador a tratar os dados de fala de forma problemática, enviesando especialmente a análise das relações sintáticas no discurso falado. Apesar disso, poucos pesquisadores atentam a este fato e se dão conta da relevância de preservar os aspectos entonacionais da fala em suas transcrições. (CRESTI apud Mittmann, em preparação).

Na escrita, todavia, segundo Moneglia (2011) é clara a identificação de unidades linguísticas maiores do que a palavra (unidades da estrutura argumental, sentenças, orações, termos nucleares e dependentes), pois a língua escrita pode ser tranquilamente segmentada de acordo com critérios sintático. Na fala, ao contrario, não é possível utilizar estes mesmos critérios para identificar unidades de referência. Evidências de corpora orais têm mostrado que aproximadamente 30% dos enunciados não apresentam um verbo e não podem ser analisadas conforme parâmetros sintáticos empregados facilmente na escrita.

Em princípio, a unidade linguística que se percebe de maneira mais natural é o turno dialógico. Entretanto, segundo Cresti (2000), o turno dialógico não pode ser considerado como unidade fundamental de referência do discurso falado, porque os turnos apresentam uma ampla variação, podendo ser compostos de apenas uma palavra ou interjeição, ou mesmo de uma longa exposição. O conceito de turno é resultado de uma interpretação antes cognitiva do que linguística.

A Teoria da Língua em Ato parte do princípio de que a unidade linguística da fala deve corresponder à unidade fundamental da atividade comunicativa, já que é esta atividade que 'sustenta' a fala. A unidade linguística fundamental da fala deve corresponder à unidade fundamental da atividade comunicativa: o ato de fala (Austin, 1962). Partindo do princípio de que a fala espontânea consiste na execução de ações, delimitar a unidade de referência da fala deve corresponder a identificar, no fluxo da fala, as sequências linguísticas que se apresentam como suficientes e autônomas do ponto de vista pragmático, isto é, as entidades linguísticas que veiculam ações. Estas unidades são identificadas com o componente linguístico, o ato locutório, do ato de fala, conforme a perspectiva de Austin (1962). Assim, o enunciado deve ser considerado como a unidade linguística básica da fala, pois corresponde ao componente linguístico de um ato de fala (CRESTI 2000, 2009a; MONEGLIA 2000, 2011; MONEGLIA; CRESTI 1993, 2006).

Esta afirmação fundamenta-se na hipótese de que seja possível estabelecer uma equivalência entre unidades do domínio das ações humanos (atos) e unidades linguísticas (enunciados). Assim, o enunciado é tido como a "contraparte linguística da ação"; isto é, o ato locutório é a contraparte linguística do ato ilocutório, e é interpretável pragmaticamente em autonomia. Isso significa, entre

outras coisas, que um enunciado não precisa necessariamente possuir um verbo, e pode, inclusive, ser composto por uma única interjeição, desde que, entoado de maneira a cumprir uma ilocução. Dessa forma, a identificação dos enunciados se realiza através de uma quebra entonacional percebida como conclusiva. Isso significa que uma unidade de enunciado (ou a única se o enunciado for simples) deve ser uma unidade de raiz (o comentário) capaz de veicular autonomia pragmática. Esse princípio baseia-se na teoria perceptiva da entonação ('t Hart; Collier; Cohen, 1990), acarretando uma relação biunívoca entre enunciado e ilocução. A cada enunciado, ou seja, a cada unidade mínima de significado pragmático, corresponde-se uma única ilocução, uma intencionalidade do falante.

Para a Teoria da Língua em Ato, as unidades informacionais são identificadas no enunciado através de três critérios distintos: o critério funcional (função exercida pela unidade no enunciado), o critério entonacional (perfil entonacional característico de cada unidade) e o critério distribucional (posição da unidade no enunciado). Dessa forma, a junção desses três critérios possibilita a identificação das unidades informacionais da fala.

Segundo a Teoria da Língua em Ato há dois tipos de unidades informacionais: as unidades textuais e as unidades dialógicas. As unidades informacionais textuais[2] são aquelas que compõem o texto do enunciado propriamente dito. Dentre elas encontramos as unidades de Comentário (COM), de Tópico (TOP), de Apêndice de Comentário (APC), de Parentético (PAR), de Introdutor Locutivo (INT), de Apêndice de Tópico (APT) e a Unidade de escansão (SCA). As unidades informacionais dialógicas ou não textuais, por sua vez, são aquelas que não contribuem para a constituição semântica de um enunciado, mas dedicam-se ao cumprimento pragmático desse enunciado sendo dirigidas ao interlocutor. São elas: Incipitário (INP), Conativo (CNT), Conector Dialógico (DCT), Fático (PHA), Alocutivo (ALL) e Expressivo (EXP).

A Unidade de Comentário é a mais importante de todas as unidades, pois é a única necessária e suficiente para execução de um enunciado. Sua função é a de realização da força ilocucionária, ou seja, a de cumprir um ato de fala. Entonacionalmente é tida como

[2] Nesse artigo serão detalhadas apenas as unidades de COM, TOP e APC, necessárias à compreensão desse estudo.

uma unidade prosódica de raiz que varia conforme o valor ilocucionário; isto é, é interpretável pragmaticamente em autonomia e possui sempre um núcleo, o qual carrega o valor funcional da ilocução. Distribucionalmente pode estar em qualquer posição no enunciado e é com relação a ela que é definida a posição das outras unidades.

A unidade de Tópico é a unidade textual cuja função é especificar no texto do enunciado o domínio de relevância ao qual a força ilocucionária se refere; isto é, o campo de aplicação da força iloucionária do comentário. Ela tem caráter opcional e é subordinada melodicamente ao comentário, não sendo interpretável autonomamente.

2.1 A unidade de Apêndice de Comentário (APC)[3]

A unidade de APC é por definição uma unidade de integração textual. A maior parte das expressões que são usadas funcionalmente como unidade de APC corresponde a um conteúdo "vazio" ou a um conteúdo genérico do ponto de vista semântico. Funcionalmente, o Apêndice integra textualmente as unidades de Comentário (COM), Comentário Ligado (COB), Comentários Múltiplos (CMM). Entonacionalmente é uma unidade tonal sem foco, com uma F0 sempre mais baixa do que a unidade da qual é apêndice, sempre com perfil nivelado ou descendente e intensidade baixa (Cresti 2000; Ulisses 2008; Oliveira 2009, 2009b, 2010). Distribucionalmente deve suceder a unidade informacional de Comentário. É tida como uma unidade de sufixo.

Ex: *REG: omitir /=COM= **só //=APC**

Estudos realizados por Tucci (2006), entretanto, com dados do C-ORAL-ROM, nos permitem distinguir entre os diferentes tipos de informação que integram a unidade de comentário. Tucci operou uma primeira classificação dessas informações em quatro categorias distintas: repetições de expressões do tema, preenchimento, retomada textual e informação tardia.

[3] Esse artigo se limita ao estudo da análise informacional da unidade de APC. As análises distribucional e entonacional não serão exploradas.

As repetições de expressões do tema do discurso são discriminadas por tipologia ou distribuição e podem ser literais (aquelas que não modificam uma dada expressão linguística) ou com variação (a repetição do conteúdo semântico apresenta-se em forma de sinônimos ou perífrases). Distribucionalmente, as repetições podem ocorrer: 1) de forma contígua: quando o conteúdo repetido é expresso no mesmo enunciado; 2) não-contígua: o conteúdo repetido é expresso em outro enunciado de um mesmo turno, ou fora de turno; e 3) *Leit Motiv:* quando ocorrem como um tipo de refrão no interior de uma conversação ou de um monólogo.

Os **preenchimentos** realizam a expansão da unidade precedente sem repetir seu conteúdo semântico ou acrescentar informações e geralmente constituem-se de advérbios ou advérbios focalizadores.

A **retomada textual** é a referência ao discurso em si ou a parte do discurso. Pode ser realizada em forma de dêixis discursiva (quando se refere ao discurso em si) ou de recontextualização (quando retoma sinteticamente uma parte do discurso);

A **informação tardia**: refere-se à adição de novas informações, quando a unidade de comentário em si é suficiente para cumprir a ilocução.

2.2 Distribuição dos APC na amostra

A unidade de APC é uma unidade informacional de tipo textual. Isto significa que a unidade de APC compõe o texto falado, diferentemente das unidades com função dialógica, cujo objetivo é regular a interação. Ainda que seja uma unidade textual, sua função é de apenas integrar a unidade de COM, já que não serve de âmbito para aplicação da força ilocucionária como a unidade de TOP, não possui função metalinguística como o PAR, nem introduz uma metailocução como o INT, tampouco possui autonomia pragmática como a unidade de COM.

A tabela 1 tem por objetivo ilustrar a proporção total de APC presentes na amostra em comparação às demais unidades textuais: tópico, apêndice de tópico, parentético e introdutor locutivo.

TABELA 1
Frequência da unidade APC em comparação às demais unidades textuais presentes na amostra (adaptada de Mittmann, em preparação)

Unidades Textuais	Quantidade	Percentual
Apêndice de Comentário (APC)	112	10%
Tópico (TOP)	604	53%
Apêndice de Tópico (APT)	23	2%
Parentético (PAR)	158	14%
Introdutor Locutivo (INT)	236	21%
TOTAL	1136	100%

Comparando as frequências relativas, a unidade de APC ocupa o quarto lugar de ocorrência no subcorpus brasileiro, com um percentual de 10% sobre o total de unidades textuais, o que nos permite afirmar não se tratar de uma unidade muito recorrente. A tabela 2 exibe uma lista com todos os textos da amostra, seguida da descrição da situação, do número de enunciados com APC, do número total de unidades terminadas e da proporção de enunciados com APC em relação ao total de enunciados de cada texto.

TABELA 2
Distribuição e proporção de enunciados com APC e descrição das situações gravadas na amostra

Texto	Situação	Enunciados com APC	Total de unidades terminadas	Proporção de enunciados com APC por enunciado
		110	5484	**2%**
Conversações		32	2039	**2%**
bfamcv01	amigos avaliam um campeonato de futebol organizado por eles e planejam o próximo	09	248	**4%**
bfamcv02	senhoras conversam sobre os preparativos do casamento de uma parente	03	385	**1%**

Continua

bfamcv03	amigos jogam sinuca	03	306	**1%**
bfamcv04	amigos jogam "Imagem e Ação", após explicar as regras do jogo para uma das participantes	03	465	**1%**
bpubcv01	funcionários de banco de sangue explicam como o sangue coletado é armazenado	05	355	**1%**
bpubcv02	reunião ordinária em uma sede regional de partido político	09	280	**3%**
Diálogos		37	2451	**1,5%**
bfamdl01	colegas de apartamento fazem as compras do mês*	05	566	**1%**
bfamdl02	colegas de faculdade batem papo enquanto organizam o material de gravação	06	282	**2%**
bfamdl03	casal faz uma viagem de carro	07	338	**2%**
bfamdl04	domésticas, mãe e filha, fazem a limpeza da cozinha após o almoço	06	253	**2%**
bfamdl05	corretor de imóveis leva a irmã para visitar apartamento	06	431	**1%**
bpubdl01	engenheiro e pedreiro trabalham em uma obra	01	276	**0,4%**
bpubdl02	cliente e vendedor interagem durante a compra de calçados	06	305	**2%**

Continua

Monólogos		41	994	**4%**
bfammn01	senhor narra história fantástica sobre uma cobra	16	106	**15%**
bfammn02	sobrinha de Carlos Drummond de Andrade conta histórias da família ao neto	06	184	**3%**
bfammn03	narrativa de "causos" divertidos para a família	04	144	**3%**
bfammn04	senhora conta sua experiência no hospital após ter dado à luz no carro*	08	189	**4%**
bfammn05	senhora fala sobre a adoção da filha após a morte de sua filha biológica	02	153	**1%**
bfammn06	pai conta seu percurso profissional à sua filha	00	76	__
bpubmn01	entrevista de avaliação sobre aulas de inglês na rede pública de ensino	05	142	**3,5%**

Quanto à proporção geral de enunciados com APC em relação ao total de enunciados de todos os textos da amostra, o resultado é de 2%, o que significa dizer que a cada 100 unidades terminadas da amostra, aproximadamente 2 deles terão APC.

A proporção relativa às conversações é de 2%, nos diálogos de 1,5%, e nos monólogos 4%. Os monólogos têm mais que o dobro de ocorrências de APC se comparado às demais tipologias.

Observando em detalhe a proporção de cada texto, repara-se que na tipologia conversação os textos bfamcv01 e bpubcv02 apresentam uma proporção de enunciados com APC maior do que o geral dessa tipologia, bem como o texto bfammn01, nos monólogos. Nos diálogos, o texto bpubdl01 apresenta proporção inferior aos

demais textos dessa tipologia. Os dois tipos de desvio podem ser explicados com base na situação comunicativa.

Quando verificadas as medidas de números de enunciados por turno, número de enunciados complexos e média de unidade tonal por enunciado, os textos bfamcv01, bpubcv02 e bpubdl01 apresentam características de um texto de maior interatividade relativamente aos outros textos de mesma tipologia. Esta característica é dada pela situação comunicativa. Já o texto bfammn01 é um texto de baixa acionalidade (que é dada pela situação comunicativa) e também possui traços de um falante com baixa diastratia. Assim sendo, pode-se afirmar que quanto maior a interatividade, menor será o número de enunciados complexos com unidades textuais, e, consequentemente, um menor número de ocorrências de APC.

Outra medida importante para a análise da amostra é a ocorrência e distribuição de unidades terminadas simples, constituídas de apenas uma unidade prosódica (enunciados simples) e unidades terminadas complexas (enunciados complexos, enunciados com unidades textuais, enunciados com unidades dialógicas, padrão ilocucionário e estrofes) compostas de mais de uma unidade prosódica e sua relação com a tipologia do texto.

Segundo Cresti (2005b), a opção por uma estrutura simples ou complexa não está ligada à formalidade dos textos, mas sim à estrutura do evento comunicativo (tipologia interacional), que opõe diálógicos (informais e formais) a monólogos (informais e formais).

É possível aprofundar o raciocínio sobre a complexidade dos enunciados, pois é importante levar em consideração diferentes tipo de enunciados complexos.

Enunciados complexos são aqueles que apresentam mais de uma unidade tonal, e, portanto, mais de uma unidade informacional – a unidade de Comentário e uma ou mais unidades textuais e/ou dialógicas. Por sua vez, os enunciados simples são aqueles que têm apenas uma unidade tonal/informacional, a unidade de Comentário. Assim, um enunciado complexo pode ser composto por:

a) comentário e uma ou mais unidades dialógicas;

b) comentário e uma ou mais unidades textuais;

c) comentário e uma ou mais unidades dialógicas em conjunto com uma ou mais unidades textuais.

Segundo Raso (2012), torna-se interessante desagregar os dados dos enunciados complexos e não tratá-los como uma única categoria, pois a natureza informacional dessa complexidade pode ter motivações diferentes nos enunciados de textos dialógicos e monológicos. Devemos tratar de maneira diferente os enunciados que são complexos porque há unidades que acrescentam informação ao conteúdo locutivo do enunciado (tipos b e c) daqueles que são complexos unicamente do ponto de vista dialógico, mas não informacional (tipo a).

De acordo com o C-ORAL-BRASIL, no grupo dialógico (diálogos e conversações) há 62% de enunciados simples, contra 38% de enunciados simples no monológico. A comparação entre enunciados complexos evidencia o oposto, isto é, 38% de enunciados simples e 62% de enunciados complexos nos monólogos.

A tabela 3 esboçará o número de unidades terminadas na amostra, excluíndo-se os enunciados simples e os enunciados complexos composto por uma ou mais unidades dialógicas.

TABELA 3
Características Informacionais do subcorpus de PB

Tipologia Informacional	Conversação	Diálogo	Monólogo
Total de unidades terminadas	**1855**	**2304**	**950**
Total de enunciados	**1534**	**1972**	**633**
Total de enunciados simples	1095	1452	351
Total de unidades escansionadas simples (COM + SCA, TMT, EMP)	91	121	63
Total de enunciados complexos com unidades dialógicas	196	232	63
Total de enunciados complexos com unidades textuais	**108**	**125**	**100**
Junção de enunciados com COM + unidades textuais e dialógicas	**44**	**42**	**56**
Total de Padrão ilocucionário	**202**	**225**	**77**
Padrão ilocucionário simples	147	148	34
Padrão ilocucionário simples com unidades escansionadas	13	19	10

Continua

Padrão ilocucionário com unidades dialógicas	24	30	8
Total de padrão ilocucionário com unidades textuais	**14**	**20**	**21**
Junção de padrão ilocuionário + unidades textuais e dialógicas	**4**	**8**	**4**
Estrofes	**119**	**107**	**240**

Esses resultados nos permitem afirmar que há mais incidência de enunciados complexos nos textos monológicos do que nos diálogos e conversações, porque os monólogos possuem unidades terminadas mais longas, além de possuírem um maior número de sequências complexas. Com base nessas características, é de se esperar a maior incidência de APC nessa tipologia textual.

Faz-se necessário esclarecer, entretanto, que o número de enunciados complexos é superior nos monólogos, considerando-se que os textos de ambas as tipologias – monólogos versus (diálogos e conversações) têm tamanhos (em número de palavras, conforme visto no capítulo 4) comparáveis.

Tendo esboçado as características informacionais do subcorpus do PB, passamos a ilustrar as unidades terminadas com unidades textuais, padrões ilocucionários e estrofes que são aquelas onde podem haver incidência da unidade de APC.

Na tabela 4 será demonstrada a proporcionalidade entre os enunciados complexos da amostra e os enunciados complexos com APC.

TABELA 4
Comparação entre o total de enunciados complexos e total de enunciados complexos com APC

Tipologias	Total de enunciados complexos c/ unidades textuais	Total de enunciados complexos com APC
Conversação	152/1855 – 8,2%	25/152 – 16,4%
Diálogo	167/2304 - 7,2%	32/167 – 19,2%
Monólogo	156/950 – 16,4%	31/156 – 20%

Como se pode notar, o número de enunciados complexos com APC na amostra é maior nas conversações e diálogos juntas (35,6%) –

pois apresentam características textuais semelhantes, isto é, o falante não tem um programa mental pré-construído, interagindo à medida que o discurso vai se construindo –, porque em ambas as tipologias, há quase o dobro de enunciados que os monólogos, conforme dito no capítulo 4.

Todavia, se desconsiderarmos essa medida veremos que são os monólogos os responsáveis pela maior incidência de enunciados complexos com APC, assim como ocorre nos enunciados complexos com unidades textuais da amostra. Esses dados corroboram o fato de que os monólogos são textual e informacionalmente mais complexos do que os diálogos.

Passemos à tabela 5 para a comparação entre o total de padrões ilocucionários da amostra e os padrões ilocucionários com APC.

TABELA 5
Comparação entre o total de padrões ilocucionários e padrões ilocucionários com APC

Tipologias	Total de padrões ilocucionários	Total de padrões ilocucionários com APC
Conversação	18/1855 – 1%	4/18 – 22,2%
Diálogo	28/2304 - 1,2 %	4/28 – 14,3%
Monólogo	25/950 – 2,6%	2/25 – 8%

Tendo por base a tabela 5, o que se verifica é que não há muitas ocorrências de padrões ilocucionários na amostra. Na verdade, sua maior incidência acontece nos monólogos, com 2,6% sobre o total analisado. Nos diálogos (conversações e diálogos) o percentual é de 2,2% sobre o total de padrões ilocucionários analisados.

Todavia, se formos analisar a incidência das unidades de APC nos padrões ilocucionários veremos que se somarmos os resultados das tipologias conversação e diálogo teremos 36,7% sobre o total de ocorrências de APC nos padrões ilocucionários, em detrimento aos 8% ocorridos nos monólogos.

Agora vejamos a incidência de APC, nas estrofes.

TABELA 6
Comparação entre o total de estrofes e estrofes com APC

Tipologias	Total de estrofes	Total de estrofes c/ APC
Conversação	119/1855 – 6,4%	2/119 – 1,7%
Diálogo	107/2304 - 4,6%	1/107 – 0,9%
Monólogo	240/950 – 25,3%	9/240 – 3,8

Como os monólogos são textos pautados principalmente no desenvolvimento semântico de um pensamento verbalizado construído por um único falante, é de se esperar que haja uma maior incidência de estrofes nessa tipologia textual (25,3%). Logo após, mas com pouca frequência estão as conversações (6,4%) e os diálogos (4,6%).

Quanto à incidência das unidades de APC nas três tipologias, o resultado é similar; isto é, há 3,8% de ocorrências de APC nos monólogos, dada a complexidade informacional dessa tipologia. Isso porque, embora o falante de um monólogo possa ter certo grau de interação com um interlocutor, sua principal atividade é a execução processual do pensamento durante a interação.

As conversações e diálogos representam juntas 2,6 % sobre o total de estrofes com APC.

As funções desempenhadas pelo APC, segundo Tucci (2006) são: repetição, preenchimento, retomada textual e informação tardia.

A tabela 7 esboça o número de ocorrências funcionais dos APC, em nossa amostra.

TABELA 7
Número de ocorrências funcionais dos APC

TEXTOS	REP	%	P	%	RT	%	IT	%	Total por texto	%
bfamcv01	-		4		3		3		10	31
bfamcv02	-		1		-		2		3	9
bfamcv03	2		-		-		1	3	9	
bfamcv04	1		-		-		1		2	6
bpubcv01	-		2		2		1		5	16
bpubcv02	1		4		-		4		9	28
	4	**12,5%**	**11**	**34,4%**	**5**	**15,6%**	**12**	**37,5%**	**32**	**100**

Continua

bfamdl01	-		2		-		3		5	14
bfamdl02	-		5		1		-		6	17
bfamdl03	2		4		-		1		7	19
bfamdl04	1		1		-		4		6	17
bfamdl05	-		2		-		3		5	14
bpubdl01	1		-		-		-		1	3
bpubdl02	-		3		1		2		6	17
	4	**11,1%**	**17**	**47,2%**	**2**	**5,5%**	**13**	**36,1%**	**36**	**100**
bfammn01	5		1		1		11		17	41
bfammn02	-		2		-		3		5	12
bfammn03	1		2		-		1		4	10
bfammn04	-		8		-		-		8	20
bfammn05	1		1				-		2	5
bfammn06	-		-		-		-		-	-
bpubmn01	-		5		-				5	12
	7	**17%**	**19**	**45%**	**1**	**2,4%**	**15**	**35,7%**	**42**	**100**
TOTAL GERAL	**15**	**14%**	**47**	**43%**	**8**	**7%**	**40**	**36%**	**110**	**100**

Os resultados apresentados na tabela 5.7 reforçam a premissa de que a unidade informacional de APC é geralmente utilizada para realizar a correção ou o acréscimo de material lexical da unidade de COM.

Na tipologia conversação, a informação tardia é a classificação informacional que mais se destaca com 37,5% de ocorrências. Em seguida estão os preenchimentos (34,4%), as retomadas textuais (15,6%), e por último as repetições (12,5%).

Nos diálogos, há a predominância dos preenchimentos (47,2%), em seguida estão as informações tardias (36,1%), repetições (11,1%) e retomadas (5,5%).

Os monólogos procedem de maneira semelhante aos diálogos. Os preenchimentos destacam-se com 45,2% de ocorrências, as informações tardias com 35,7%, as repetições com 16,6% e as retomadas textuais com 2,4%.

Em âmbito geral, as unidades de APC no PB desempenham mais a função de preenchimento, com 43% de ocorrências sobre o total de APC, seguido pela informação tardia com 36%, depois as repetições com 14% e, por último, as retomadas textuais incidindo em apenas 7% sobre o total de APC na amostra.

Por se tratar de uma classificação bastante coerente, optamos apenas por tentar aproximar as categorias propostas por Tucci (2006)

do PB, propondo uma subdivisão de duas categorias: as repetições e as informações tardias.

Segundo Tucci (2006), as repetições de expressões do tema do discurso são discriminadas por tipologia ou distribuição, podendo ser literais ou parciais. Distribucionalmente, as repetições podem ocorrer de forma contígua, não-contígua e *Leit Motiv*.[4]

Nossa proposta concentra-se na inclusão de mais uma discriminação por tipologia, a repetição parcial de forma contígua ou não. Vejamos os exemplos.

Exemplo 1 (bpubcv02)
*CAR: [212] eu nunca fui na casa lá dela não /=CMM= mas diz que é <muito ruim> /=CMM= **lá o local** //=APC=$

Exemplo 2 (bfamdl03)
*LUZ: [1] porque /=DCT= eu só soube que eu nu +=EMP=$ [2] eu tive certeza absoluta que eu nu era daqui quando eu saí //=COM=$ [3] que eu senti que então /=SCA= eu tava no meu lugar /=COM= né //=PHA=$ [4] porque eu [/1]=SCA= eu me senti /=SCA= respirando /=i-COB= né /=PHA= adequada /=COM= né /=PHA= **no lugar** //=APC=$ [5]

Exemplo 3 (bfammn01)
*MAI: [20] é um trilho dentro da mata /=COB= a mata virgem //=COM=$ [21] n' é matinha igual essas capoeirinha aqui não /=COB= é mata mesmo /=COB= de /=SCA= madeira /=SCA= da grossura que /=SCA= quato homem nu abarca um pau //=COM=$ [22] uma /=SCA= tora /=COM= **da madeira** //=APC=$

Nos três exemplos listados acima o que se viu na unidade de APC foi a reformulação de uma informação dada pelo contexto e não apenas uma repetição não-literal dessa informação, embora deva-se admitir que, às vezes, essa informação é quase insignificante, quando tomada apenas semanticamente.

Proposta a reclassificação da função informacional repetição, passamos a esboçar na tabela 8 o número de ocorrências de APC, com diferentes funções de repetição, na amostra.

[4] Em nossa amostra não foi encontrado nenhum APC com função informacional de *Leit Motiv.*

TABELA 8
Número de ocorrências de APC com diferentes funções de repetição

Textos	Total de ocorrências de APC com função de repetição	Repetição Literal		Repetição Parcial		Repetição com Variação
		Contígua	**Não-contígua**	**Contígua**	**Não-contígua**	
bfamcv01	-	-	-	-	-	-
bfamcv02	-	-	-	-	-	-
bfamcv03	2 – 50%	-	1	1	-	-
bfamcv04	-	-	1	-	-	-
bpubcv01	1- 25%	-	-	-	-	-
bpubcv02	1- 25%	-	-	-	-	1
Total Parcial	**4- 100%**	**-**	**2 – 13%**	**1 –7%**	**-**	**1 – 7%**
bfamdl01	-	-	-	-	-	-
bfamdl02	-	-	-	-	-	-
bfamdl03	2 – 50%	-	1	-	-	1
bfamdl04	1 – 25%	-	-	-	1	-
bfamdl05	-	-	-	-	-	-
bpubdl01	1 – 25%	-	1	-	-	-
bpubdl02	-	-	-	-	-	-
Total Parcial	**4 – 100%**	**-**	**2 – 13%**	**-**	**1 – 7%**	**1 – 7%**
bfammn01	5- 72%	2	-	1	1	1
bfammn02	-	-	-	-	-	-
bfammn03	1- 14%	-	-	-	1	-
bfammn04	-	-	-	-	-	-
bfammn05	1-14%	-	1 -	-	-	-
bfammn06	-	-	-	-	-	-
bpubmn01	-	-	-	-	-	-
Total Parcial	**7 – 100%**	**2 – 13%**	**1 – 7%**	**1 – 7%**	**2 – 13%**	**1 – 7%**
TOTAL GERAL	**15- 100%**	**2 – 13%**	**5 – 33%**	**2 – 13%**	**3 – 20%**	**3 – 21%**

Embora os percentuais de ocorrência desse tipo de classificação informacional seja muito pequeno, podemos dizer que as repetições literais são as mais comuns entre as possíveis funções do APC, com um percentual de 46% sobre o total de APC. Em seguida, estão as repetições parciais com 33% de ocorrências e, em último, as repetições com variação (21%).

Quanto ao fato de as repetições ocorrerem dentro ou fora do enunciado, achamos ser desnecessário pontuar o percentual, já que se tratam de valores pouco representativos.

As informações tardias, segundo Tucci (2006), referem-se à adição de novas informações à unidade de COM. Após análise de nossos dados, vimos que a informação tardia é caracterizada também pela composicionalidade sintática com o conteúdo expresso na unidade de COM. Faz-se necessário esclarecer, entretanto, que nem sempre é possível interpretar se o APC compõe sintaticamente com a unidade de COM, mas não se dá nenhum caso de informação tardia sem que seja possível a composicionalidade sintática. Analisemos alguns exemplos para melhor ilustrar a afirmação acima.

Exemplo 4 (bfamcv01)
*GIL: [65] da gente /=TOP= pra alguns setores da organização /=TOP= chamar o pessoal /=COM= dos outro times //=APC=$

A unidade de APC nesse caso integra a informação da unidade de COM, procurando elucidar de que 'pessoal' se trata (o pessoal 'dos outro times'). Veja-se que a unidade de APC passa a compor sintaticamente com a unidade de COM ('chamar o pessoa'. Que pessoal? 'O dos outros times'). Agora, passemos a outro exemplo.

Exemplo 5 (bfamcv01)
*GIL: [137] <é> //=COM=$ [138] todo mundo <encheu o saco> /=COM= <**por causa disso**> //=APC=$

Nesse exemplo, o APC pode ou não ser interpretado com composicionalidade sintática.

Dessa forma, podemos dizer que se o APC só puder ser interpretado por meio da composicionalidade sintática, então será obrigatoriamente classificado como sendo uma informação tardia.

Na amostra, há, também, um único exemplo de APC com função de informação tardia e valor de glosa. Vejamos.

Exemplo 6 (bfammn01)
MAI: [9] e quando chegou lá /=TOP= &he /=TMT= montou uma casinha pa ele /=COB= pa família dele /=COB= e tal /=COB= e e' vinha na cidade pa compar alguma coisa &dif [/1]=SCA= diferente /=COM= **que nu era da roça /=APC**= né //=PHA=$

Repare que essa glosa visa a reforçar o que fora mencionado na unidade de COM, funcionando como uma espécie de nota explicativa, já que se diz 'compar alguma coisa diferente', diferente de quê? De 'que nu era da roça'.

A retomada textual é a referência ao discurso em si ou a parte do discurso. Pode ser realizada em forma de deixis discursiva, quando se refere ao discurso em si, ou de recontextualização, ao retomar sinteticamente uma parte do discurso. A título de exemplo, ilustramos alguns casos.

Exemplo 7 (bfamcv01)
*EVN: [151] é /=INP= a <gente tem que> <restringir também /=COM= **isso**> //=APC=$

Nesse caso, a unidade de APC, por meio do pronome 'isso', é um encapsulador, cuja função é retomar algo mencionado.

Agora vejamos um exemplo de retomada com reformulação.

Exemplo 8 (bfammn01)
*MAI: [85] na Amazonas tem muito dessa cobra aí //=COM=$[86] só ni lugar /=SCA= que tem /1 =SCA= tem mata mata muito /=SCA= forte /=COB= mata /=SCA= perigosa /=COB= que tem <esses tipo> de cobra /=COM= né //=PHA=$
*DUD: [87] <uhn> //=COM=$
*MAI: [88] no norte de Mina /=TOP= tinha esse /2 =SCA= antigamente /=PAR= tinha esse tipo de cobra tudo /=COM= né //=PHA=$[89] talvez agora já acabou /=COB= porque já desmataram muito /=COM= né //=PHA=$[90] nu existe muita mata mais igual /1 =SCA= igual era /=COB= mas tem muita mata forte aí /=COB=

nu é igual essas /1 =SCA= essas capoeirim /=COB= cê fala assim /=INT= ah /=EXP_r= isso aqui é uma mata //=COM_r=$[91] isso aqui /=TOP= em vista de lá /=TOP= é /1 =EMP= é um cerrado /=COM= **aqueas mata lá //=APC**=$

Nesse exemplo, a unidade de APC funciona como paráfrase daquilo que fora mencionado no enunciado 85.

Os APC com função de preenchimento foram os mais encontrados na amostra. Eles representam 43% sobre o total de APC analisados e tem como objetivo realizar a expansão da unidade precedente sem repetir seu conteúdo semântico ou acrescentar informações.

Exemplo 9 (bfammn01)
MAI: [79] é lugar /=SCA= perigoso mesmo o norte de Minas /=COM= lá //=APC=$

Exemplo 10 (bfammn03)
*ALO: [37] sio' /=SCA= precisa de ir lá p' siora despedir //=COM_r=$ [38] do pai lá /=COM_r= **e tal //=APC**_r=$

Embora tenhamos ampliado a proposta de Tucci, ainda assim há dois casos que nos deixaram muito inseguros quanto à sua classificação informacional. Vejamos:

Exemplo 11 (bfamdl05)
*ANE: [207] <só tem esse> apartamento pra vender //=COM=$
*CES: [208] obrigado> //=COM=$
*ENC: [209] eu creio que sim //=COM=$ [210] que +=EMP=$
*ANE: [211] tá //=COM=$
*ENC: [212] tá //=COM=$
*ANE: [213] esquecemo de perguntar <pa ele a> previsão /=COM=né //=PHA=$
*CES: [214] <bom> //=COM=$ [215] é //=COM=$ [216] deixa eu <perguntar pra ele /=COB= a> previsão //=COM=$
*ANE: [217] <pergunta lá> //=COM=$
*CES: [218] eu acho que a previsão dele tá pra entregar em dezembro //=COM=$ [219] a previsão de entrega do prédio //=COM=$

*ENC: [220] no começo yyyy //=COM=$
*CES: [221] ahn //=COM=$
*ENC: [222] no começo yyyy //=COM=$
*CES: [223] então /=INP= até final de novembro //=COM=$
*ENC: [224] isso //=COM=$
*ANE: [225] &he /=TMT= essa localização aqui /=TOP= eu nu acho ruim //=COM=$
*RAQ: [226] eu gostei também foi por causa do construído já ao redor /=COM= <né> //=PHA=$
*ANE: [227] <é> //=COM=$
*RAQ: [228] <é sim> //=COM=$
*ANE: [229] <tá tudo> <construído> //=COM=$
*CES: [230] <Anete> //=COM=$ [231] <ele> [/1]=EMP= e' disse que é &d +=EMP=$ [232] final de novembro //=COM=$ [233] <é /=INP= porque agora é rápido> /=COM= **isso <aqui**> //=APC=$

Em princípio, havíamos classificado o APC acima como sendo uma retomada textual, porque acreditávamos que o pronome 'isso' referia-se à entrega do prédio. Todavia, ao fazermos a revisão de toda a classificação vimos que esse APC também poderia ser classificado como preenchimento, já que é semanticamente vazio.

Outro caso que nos deixou indecisos foi um enunciado retirado do texto bfammn02.

Exemplo 12 (bfammn02)
DFL: [165] e às vezes ele tava lendo jornal /=TOP= de vez em quando ele [/1]=SCA= ele interrompia /=COB= ele tava sempre atento às nossas brincadeira /=PAR= e' falava /=INT= o que é o que é //=COM_r=$ [166] a gente <assustava /=COB= e tinha que +=EMP=$ [167] às vezes +=EMP=$ [168] se sabia /=CMM= falava /=CMM= <mas se nu sabia /=CMM= né> //=PHA=$
*LUC: [169] <uma charada /=COB= **ali**> /=APC= <na hora> //=COM=$

Parece que esse APC, a priori, seria uma retomada textual, desde que se considerasse que o advérbio 'ali' refere-se ao discurso em si, ao fato de 'ali' coincidir com o tempo em que ocorriam as brincadeiras; mas se não interpretarmos dessa maneira, poderíamos dizer que se trata de uma informação tardia, algo inserido entre as

unidades de COM apenas para reafirmar o lugar puramente discursivo em que ocorriam as brincadeiras.

Embora tenhamos dúvidas quanto à classificação informacional da unidade de APC em apenas dois casos, pensamos poder afirmar que nem sempre a unidade de APC apresenta uma única função; pode, por vezes, ser interpretada de maneiras distintas dado o contexto enunciativo.

Referências

AUSTIN, J. *How to do things with words*. Oxford: Clarendon. Bernardo (a c. di), Logica deontica e semantica, Bologna, Il Mulino, 1962. p. 147-165. Bernard Pottier. Représentations mentales et catégorisations linguistiques, Louvain e Paris, Peeters, 2000.

CRESTI, E. The definition of Focus in Language into Act Theory. V° LABLITA WORKSHOP & II° BRAZILIAN SEMINAR ON PRAGMATICS AND PROSODY: ILLOCUTION, MODALITY, ATTITUDE, INFORMATION PATTERNING AND SPEECH ANNOTATION. *Anais...* Belo Horizonte, 2010.

CRESTI, E. La stanza: un´unità di costruzione testuale del parlato. In: X CONGRESSO SILFI: SINTASSI STORICA E SINCRONICA DELL´ITALIANO. SUBORDINAZIONE, COORDINAZIONE, GIUSTAPPOSIZIONE. *Atti...* Firenze: Cesati, 2009.

CRESTI, E. Enunciato e frase: teoria e verifiche empiriche. In: BIFFI, M.; CALABRESE, O.; SALIBRA, L. (Org.). *Italia Linguistica*: discorsi di scritto e di parlato. Scritti in onore di Giovanni Nencioni. Siena: Prolagon, 2005a. p. 249-260.

CRESTI, E. Corpus di italiano parlato. Firenze: Accademia della Crusca, 2 voll. 2000.

CRESTI, E.; MONEGLIA, M. Informational patterning theory and the corpus-based description of spoken language: The compositionality issue in the topic-comment pattern. In: PANUNZI, A. (Ed.). *Bootstrapping Information from Corpora in a Cross- Linguistic Perspective*. Firenze: FUP, 2010.

CRESTI, E.; MONEGLIA, M. C-ORAL-ROM. Comparing Romance Languages in Spontaneous Speech Corpora. In: SILVA, T. C.; MELLO, H. R. (Ed.). V CONGRESSO INTERNACIONAL DA ASSOCIAÇÃO BRASILEIRA DE LINGUÍSTICA. *Anais*... Belo Horizonte: UFMG, 2007.

't HART, J.; COHEN, A.; COLLIER, R. *A perceptual study on intonation*: an experimental approach to speech Mellody. Cambridge: Cambridge University Press. Laboratorio Linguistico del Dipartimento di Italianistica dell Università di Firenze. 1990. Disponível em: <http://lablita.dit.unifi.it/>. Acesso em: 10 julho 2007.

MARCUSCHI, L. A. *Da fala para a escrita*: atividades de retextualização. Sao Paulo: Cortez. 2001.

MITTMANN, M. *A unidade informacional Topico na fala espontanea*: um estudo baseado em corpus. Tese (Doutorado em Linguistica) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

MONEGLIA, M. Spoken corpora and pragmaticis. *Revista Brasileira de Linguistica Aplicada*. Belo Horizonte, v. 11, n. 2, p.479-519, 2011.

MONEGLIA, M. Specifications on the C-ORAL-ROM Corpus. 2000. Disponível em: http://lablita.dit.unifi.it/coralrom/papers/Specifications-CORALROM.pdf

MONEGLIA, M.; CRESTI, E. C-ORAL-ROM. Prosodic boundaries for spontaneous speech analysis. In: KAWAGUCHI, Y.; ZAIMA, S.; TAKAGAKI, T. (Ed.). *Spoken Language Corpus and Linguistics Informatics.* Amsterdam: John Benjamins, 2006.

OLIVEIRA, C. Apêndice de Comentário e Comentários Ligados: uma distinção à luz da Teoria da Língua em Ato. *Revista Eletrônica de Divulgação Científica em Língua Portuguesa, Linguística e Literatura Letra Magna*. Ano 05, n. 10, 1º semestre, 2009. Disponível em: WWW.letramagna.com.

OLIVEIRA, C. Distinção entre Apêndice de Comentário, Comentários Ligados e Inciso, três unidades informacionais, em final de enunciado, à luz da Teoria da Língua em Ato. *Revista e_Hum*, v. 2. n. 1, 2009b.

OLIVEIRA, C. O Apêndice de Comentário: classificação informacional e morfossintática à luz da Teoria da Língua em Ato. *Revista Eletrônica de Divulgação Científica em Língua Portuguesa, Linguística e Literatura Letra Magna*. Ano 06, n. 13, 2º semestre, 2010. Disponível em: WWW.letramagna.com.

RASO, T.; MELLO, H. (Org.). C-ORAL-BRASIL. *Corpus de referência da fala espontânea informal do português do Brasil*. Belo Horizonte: UFMG, 2012.

TUCCI, E. L'unità di appendice in un corpus di italiano parlato (C-ORAL-ROM): caracteristiche intonative, semantiche e morfo-sintattiche. Tesi de laurea triennale in italianistica. Universitá degli studi di Firenze, Facoltà di lettere e filosofia, anno academico. 2005/2006.

ULISSES, Andréa de Jesus. *Unidade de apêndice no português do Brasil*. Dissertação (Mestrado) - Faculdade de Letras da UFMG. 2008.

Um estudo do Tema Predicado no português brasileiro: contribuições de uma abordagem de *corpus*

Giacomo Figueredo[1]
Adriana Pagano[2]
Kícila Ferreguetti[3]

RESUMO: Este trabalho apresenta resultados preliminares de uma pesquisa em andamento, realizada com subsídios metodológicos da Linguística de *Corpus*. Mais especificamente, utiliza uma abordagem de *corpus* para investigar o Tema Predicado – um dos recursos gramaticais característicos que responde por realizar a focalização na oração (HALLIDAY, 2005). Do ponto de vista discursivo, o Tema predicado confere proeminência textual a elementos que, ao serem predicados, contribuem para mudar o fluxo discursivo e redirecioná-lo conforme o tipo de texto (MATTHIESSEN, 1995). Uma vez que este recurso é utilizado em diferentes tipos de texto, uma abordagem de *corpus* pode contribuir para sua descrição aliada ao seu uso. Este trabalho propõe assim uma investigação do Tema Predicado a partir da observação de ocorrências extraidas do *corpus* CALIBRA, compilado a partir de textos representativos da língua portuguesa do Brasil. O corpus CALIBRA (Catálogo da Língua Brasileira), desenvolvido na FALE/UFMG e DELET/UFOP, possui aproximadamente um milhão de palavras e contempla textos orais e escritos pertencentes a diferente tipos textuais. A análise das ocorrências extraídas do *corpus* é feita dentro do âmbito da descrição

[1] Doutor em Linguística Aplicada. Professor Adjunto de Linguística Aplicada. Universidade Federal de Ouro Preto. giacomojakob@yahoo.ca.

[2] Doutora em Letras. Professora Associada da Universidade Federal de Minas Gerais. apagano@ufmg.br.

[3] Mestranda em Estudos Linguísticos (Estudos da Tradução). Universidade Federal de Minas Gerais. kicilaferreguetti@yahoo.com.br.

sistêmico-funcional, em sua relação com o sistema de TEMA no português brasileiro (FIGUEREDO, 2011).

PALAVRAS-CHAVE: Corpus Monolíngue, Teoria Sistêmico-Funcional, Tema Predicado, Descrição Linguística, Português Brasileiro.

ABSTRACT: This paper reports on an ongoing corpus-based research of Theme Predication in Brazilian Portuguese. Theme Predication is one of the resources within the textual grammatical systems of Brazilian Portuguese that have developed to give prominence status to a given element in the clause. The resources of Theme Predication, in particular, are responsible for realizing grammatically the meanings related to focus, contrast and definition (BRAGA, 2009, p. 178; LONGHIN e ILARI, 2000, p. 203). From a discourse point of view, Theme Predication is complementary deployed at specific points in the text, contributing to reorient the flow towards the field (MARTIN e ROSE, 2007, p. 188). Aiming at investigating Theme Predication at clause level, as well as its role in discourse, this paper draws on corpus data and adopts a systemic functional view on description of grammatical behavior as well as use in discourse of this function.

KEYWORDS: Monolingual Corpus, Systemic Functional Theory, Predicated Theme, Brazilian Portuguese, Language Description

1 Introdução

Este trabalho apresenta resultados preliminares de uma pesquisa em andamento, realizada com subsídios metodológicos da Linguística de *Corpus*, no âmbito de projetos conjuntos do Grupo de Pesquisa Produção de Significado em Ambientes Multilíngues (DELET/UFOP) e do Laboratório Experimental de Tradução (FALE/UFMG). Mais especificamente, utiliza uma abordagem de *corpus* para investigar as funções gramaticais que realizam os significados de focalização em português brasileiro sob uma perspectiva sistêmico-funcional. A partir de uma abordagem de *corpus*, busca-se examinar as funções relativas ao Tema Predicado em português brasileiro, tendo como objetivo explicitar as contribuições que tal abordagem pode oferecer tanto para a descrição gramatical do Tema Predicado quanto para a observação do seu emprego no fluxo discursivo.

Toda pesquisa sobre algum aspecto da linguagem em particular, quando conduzida por uma abordagem de *corpus* – utilizando, portanto, subsídios metodológicos da Linguística de *Corpus* – pressupõe, em primeiro lugar, a utilização de dados empíricos extraídos segundo uma metodologia objetiva a partir de um conjunto de textos cuja representatividade seja compatível com o objetivo do estudo (VIANA, 2011, p. 28). Em segundo lugar, os padrões linguísticos encontrados (que, no caso específico deste trabalho, visam à descrição de determinada função linguística) são, necessariamente, interpretados a partir do que se observa no *corpus*, conformando os pressupostos da teoria aos dados efetivamente extraídos do *corpus*. A partir desse ponto, no presente trabalho, faz-se o cotejo entre os padrões linguísticos extraídos do *corpus* com os pressupostos da teoria linguística para, assim, propor uma descrição do comportamento e do uso, incluindo distribuição e frequência ao longo dos textos e dos tipos de texto, da função do Tema Predicado.

No que tange especificamente à função de Tema Predicado dentro das formulações da teoria sistêmico-funcional, esta se relaciona à organização textual do sistema linguístico; em particular à tessitura estrutural (HALLIDAY e HASAN, 1976, p. 325). A tessitura é uma propriedade fundamental da língua, uma vez que cabe a ela conferir a natureza simbólica (ou "semiotização"; MATTHIESSEN, 1995, p. 21) para as situações sociais e materiais das quais nós, falantes, tomamos parte.

Isto quer dizer que, em uma determinada situação comunicativa, os falantes tendem a produzir os significados relevantes para essa situação. Estes significados podem ser de natureza interpessoal, os quais estabelecem e mantêm a relação entre os interlocutores (por exemplo, hierarquia, polidez, avaliação, entre outros); ou de natureza experiencial, os quais representam a experiência de mundo dos interlocutores (por exemplo, a organização das ações, acontecimentos, relações lógicas, entre outros). Cabe à tessitura, então, converter o conjunto de significados linguísticos produzidos para uma determinada situação (de natureza interpessoal ou experiencial) em uma unidade semântica, à qual se denomina 'texto'.

A unificação imposta pela tessitura à situação (i.e., organização social e material) – cujo produto é o texto – é modelada pela teoria linguística como um movimento de ondas de informação (HALLIDAY, 2002, p. 209). Estruturalmente, a tessitura pode oferecer proeminência

de informação a determinados elementos, para assim orientarem o fluxo discursivo; e não oferecer proeminência para outros, para assim tomarem parte no fluxo previamente estabelecido (MATTHIESSEN, 1992, p. 41).

Para o português brasileiro, assim como para outras línguas (CAFFAREL *et al.*, 2004, p. 53), o sistema principal da organização estrutural é o TEMA. Este é um sistema da oração, e as funções geradas por ele selecionam elementos gramaticais interpessoais e ideacionais para desempenhar dois papéis fundamentais: (i) conectar a oração ao discurso precedente; e (ii) nos limites da oração, produzir a base de interpretação para o restante da oração.

Ao conectar a oração ao discurso precedente, o TEMA faz com que esta tome parte no fluxo discursivo (conferindo-lhe, ou não, proeminência), incluindo-a, desta forma, em um tipo de texto específico. É desta maneira que a tessitura, realizada gramaticalmente pelas funções temáticas promove, a cada escolha nas orações, a unificação semântica do texto – ou sua "semiotização" (HALLIDAY e MATTHIESSEN, 2004, p. 64). Dentro dos limites da oração, a função temática manipula os elementos oracionais de forma a criar um micro-contexto de interpretação, relativamente à proeminência/não-proeminência destes elementos (MATTHIESSEN, 1995, p. 531).

A consolidação das pesquisas acerca da forma como o TEMA organiza o fluxo discursivo tem sido capaz de revelar diferentes aspectos da sua constituição e forma de funcionamento. Do ponto de vista do discurso, é possível compreender como os textos da língua se distribuem em tipos (registros e gêneros) segundo a sua organização textual (HERKE-COUCHMAN, 2006; MARTIN e ROSE, 2007). Do ponto de vista da oração, compreende-se como os recursos gramaticais se dispõem (HALLIDAY e MATTHIESSEN, 2004) e a forma como cada um deles é capaz de criar o micro-contexto oracional.

No caso específico do Tema Predicado, é possível compreender o seu papel na criação de foco e contraste no decorrer do fluxo do discurso (MATTHIESSEN, 1995, p. 554; HALLIDAY e MATTHIESSEN, 2004, p. 95). Em português brasileiro, este recurso é bastante estudado, e seu funcionamento bastante conhecido, em particular nas pesquisas sobre clivagem e pseudo-clivagem, das construções 'SER...QUE', 'QUE', 'É QUE', dos quais se destacam as pesquisas de Braga (1991; 2009) e, especificamente com relação à organização gramatical sistêmica, Longhin e Ilari (2000).

Tomando a descrição de Halliday para a língua inglesa, Longhin e Ilari (2000), investigam em que medida a descrição dos sistemas textuais de TEMA, IDENTIFICAÇÃO e INFORMAÇÃO se aplica ao português. Na tentativa de apresentar os diferentes recursos de clivagem (i.e., gramaticalização da focalização), esses autores mostram propriedades das construções de identificação que permitem explorar os recursos de produção de significado dessa língua, dentre as quais ganha destaque o Tema Predicado.

O avanço nesta compreensão, por conseguinte, oferece as bases para um estudo a partir de uma abordagem de *corpus*, que pode complementar o conhecimento acerca dos sistemas textuais com dados extraídos dos diferentes tipos de texto, bem como elucidar as diferenças de uso com base na configuração de cada tipo de texto. Desta forma, o presente trabalho visa contribuir com as pesquisas sobre a constituição sistêmica da gramática que realiza a tessitura. Mais especificamente, investigando um dos recursos importantes do sistema de TEMA – o Tema Predicado – a partir das contribuições de uma abordagem de *corpus*. Assim, é objetivo específico da presente pesquisa – que está em andamento, e da qual este texto apresenta resultados preliminares – analisar o Tema Predicado, observando a partir do *corpus* a sua constituição, frequência e ocorrência por tipo de texto e ao longo dos textos – tendo como base a sua concepção sistêmico-funcional – a manipulação do micro-contexto oracional com objetivo de focalizar determinado elemento para re-orientar o fluxo discursivo.

2 O "endereço semiótico" do Tema Predicado no espaço do sistema linguístico

A modelagem da língua como sistema – o sistema linguístico – implica em compreender, em primeiro lugar, como este se organiza internamente (HALLIDAY e MATTHIESSEN, 2004, p. 24). Somente a partir da compreensão desta organização interna é possível investigar a relação entre o sistema linguístico e o meio no qual é produzido (i.e. a situação relevante para a comunicação). Assim, na modelagem sistêmica, é papel do sistema linguístico "metaorganizar" a situação linguisticamente, conferindo a esta a condição de símbolo linguístico, ou significado. As funções internas do sistema linguístico, responsáveis por esta metaorganização são assim denominadas metafunções.

A metafunção ideacional é responsável por representar a experiência do mundo e possui como sistemas gramaticais principais a AGÊNCIA e a TRANSITIVIDADE (NUCLEAR e CIRCUNSTANCIAL) na ordem da oração. A metafunção interpessoal é responsável por encenar a interação social e tem como sistemas principais da gramática o MODO e a MODALIDADE. A metafunção textual é responsável por habilitar a realidade semiótica organiza as outras metafunções, contextualizando o texto na situação, dentro de um tipo de texto determinado.

Os sistemas gramaticais da metafunção textual são responsáveis pela materialização linguística da tessitura estrutrural – o TEMA e seus sub-sistemas – e, desta forma, distribuem as funções no fluxo discursivo. Estes confluem com um sistema fonológico, a INFORMAÇÃO, que apresenta a informação, conferindo a um elemento (a confluência entre função gramatical e unidade fonológica) a proeminência ou a não-proeminência no fluxo do discurso.

A INFORMAÇÃO é um sistema prosódico-entonacional que responde pela apresentação da informação (HALLIDAY e GREAVES, 2008; CAGLIARI, 1981), dividindo-a entre informação que dá continuidade ao fluxo discursivo (realizada pela função de Dado) e não recebe proeminência fonológica; e informação que promove a descontinuidade no fluxo (realizada pela função de Novo) e recebe proeminência fonológica. O sistema de TEMA se constitui como o recurso empregado para a manipulação da contextualização oracional (MATTHIESSEN, 1995, p. 531). Do ponto de vista do discurso, o TEMA seleciona uma função ideacional e outra interpessoal e promove sua confluência para que operem como ponto de partida para a interpretação da oração, colocando-a como parte do fluxo discursivo, consoante com o tipo de texto. O Tema Predicado é uma opção mais delicada do sistema de TEMA, que gera o sub-sistema denominado TEMA PREDICADO, cuja função principal é colocar o Novo na posição temática com função não-marcada (CAFFAREL, 2006; MATTHIESSEN, 1995). Assim, dentre as diferentes estratégias que a gramática textual das línguas pode empregar, o Tema Predicado desempenha um papel de destaque. Funcionalmente, este tipo de Tema responde, no nível da oração, pela explicitação dos papéis experiencial e interpessoal do elemento predicado (MATTHIESSEN, 1995, p. 554); no discurso, responde pela mudança na fase discursiva.

Desta maneira, qualquer elemento da oração que realize uma função experiencial, juntamente com uma função interpessoal, pode

ganhar proeminência temática através dos recursos do Tema Predicado, dependendo do tipo de texto e do momento do fluxo discursivo. Por exemplo (todos os exemplos apresentados foram retirados do *corpus* desta pesquisa; cf. Seção 3, a seguir):

Exemplo 1 - *Os cálculos aproximados revelaram uma população de 16.101 indivíduos, ou seja, 500 vezes mais do que o número de baleias avistadas. São usadas fórmulas matemáticas para calcular o número total de baleias a partir de números menores resultantes das avistagens reais. Essas fórmulas tentam levar em conta diversas variáveis, tais como o fato de que certas espécies de baleias tendem a crescer juntas em certas áreas.* ***E é nessas fórmulas*** *que existe um potencial enorme para erros acumulados.*

Exemplo 2 - *A mesma supervalorização irradia-se pelo campo psíquico e se manifesta como uma cegueira lógica (enfraquecimento do juízo) perante as realizações anímicas e as perfeições do objeto sexual, e também como uma submissão crédula aos juízos dele provenientes.* ***Assim é*** *que a credulidade do amor passa a ser uma fonte importante, se não a fonte originária da autoridade.*

Exemplo 3 - ***Ora, é essa supervalorização sexual*** *que não suporta bem a restrição do alvo sexual à união dos órgãos genitais propriamente ditos e que contribui para elevar as atividades ligadas a outras partes do corpo à condição de alvos sexuais.*

Exemplo 4 - *Eu era um agricultor da Serra de Escambray, no centro de Cuba.* ***Foi ali*** *que ouvi falar de revolução pela primeira vez.*

Nestes exemplos, verifica-se a forma como a gramática textual, por meio da função do Tema Predicado, opera a confluência entre uma função experiencial e outra interpessoal, dando proeminência gramatical a este elemento na posição temática. No exemplo 1, "nessas fórmulas" a confluência experiencial/interpessoal é feita por Circunstância:Localização /Adjunto. Em 2, Circunstância:Modo/ Adjunto; em 3, Ator/Sujeito. Finalmente, no exemplo 4, Circunstância: Localização /Adjunto.

O significado de foco (proeminência, contraste ou particularização) se dá pela singularização gramatical do elemento predicado, nas orações que sempre se configuram como uma igualdade entre este elemento e sua definição ou descrição. No caso do exemplo 1, há uma igualdade entre o local "nessas fórmulas" e o "potencial

enorme". No exemplo 2 a maneira "assim" é igualada à "credulidade do amor passar a ser..."; em 3, "essa supervalorização sexual" e o fato de ela "não suportar bem a restrição...". No exemplo 4, o local "ali" se iguala ao local onde o agricultor "ouviu falar de revolução...".

No que diz respeito à materialização linguística – i.e., a realização gramatical – como se pode observar, na estrutura do Tema Predicado está, no primeiro termo, a configuração "ser ^ Elemento (que)", "Elemento ^ ser (que)", ou ainda apenas "Elemento (que)" e, no segundo termo, o encaixamento de uma oração ao elemento destacado no Tema, sempre elaborando-o (na fórmula que existe um potencial enorme; a maneira como a credulidade do amor passa a ser; a supervalorização que não suporta bem a restrição; o local onde ouviu falar da revolução).

A partir destas considerações sobre a teoria e ambiente descritivo da função do Tema Predicado, passa-se, na seção seguinte, à metodologia que leva à análise dos dados.

3 Descrição gramatical orientada por uma abordagem de *corpus*

Viana (2011) sintetiza os pressupostos sobre o *corpus* pelos quais se pauta esta pesquisa da seguinte forma:

> Uma compilação eletronica e criteriosa de (amostras de) textos que ocorrem naturalmente com o objetivo de representar uma dada língua ou algum de seus aspectos mais pontuais de forma a possibilitar uma análise linguística previamente delineada. De outra forma, isso significa afirmar que um *corpus*: a) deve ser compreendido como um conjunto de textos; b) contempla textos (orais e escritos) que tenham sido efetivamente produzidos por falantes de determinada língua; c) consiste numa forma de representar empiricamente o uso que se faz de uma língua em seu sentido geral ou específico; d) é uma reprodução da produção linguística de toda a população que se quer investigar ou uma amostra representativa dessa população, com base em princípios claros e bem definidos; e) assume a forma eletrônica com vistas a ser investigado pelo computador; f) é concebido com o objetivo de possibilitar a realização de uma pesquisa linguística (VIANA, 2011, p. 27).

Do ponto de vista teórico, buscou-se o respaldo metodológico nos princípios de descrição da lingüística sistêmico-funcional, a qual propõe adotar diferentes pontos de vista analítico-descritivos – a visão trinocular (Halliday, 2002, p. 408). Esta permite triangular três pontos analíticos acerca dos fenômenos, investigando-os: (I) "de baixo" (da manifestação para a organização gramatical), observando como são realizados pelos elementos constituintes da estrutura; (II) "ao redor" (nas relações funcionais), descrevendo as opções sistêmicas (oposição, delicadeza e *valeur*); (III) "de cima" (do significado para a organização gramatical), desde a semântica, examinando os significados produzidos pelas funções gramaticais no desenvolvimento do texto.

Desta maneira, a adoção da visão trinocular permite orientar a pesquisa de forma a buscar responder às seguintes perguntas:

1) DE BAIXO: quais os diferentes tipos de realização para as funções do Tema Predicado? Em quais estruturas estas funções se manifestam? Em que medida as diferenças estruturais realizam funções diferentes?
2) AO REDOR: como as funções do Tema Predicado se relacionam com os sistemas informacionais, TEMA e INFORMAÇÃO?
3) DE CIMA: em que momento do desenvolvimento do fluxo discursivo se emprega a focalização? Existem restrições no seu emprego por tipo de texto? Como esta contribui para o significado do texto?

4 Compilação do *corpus*

A partir dos pressupostos da Linguística de *Corpus*, a compilação dos textos que formam a base analítica deve ser representativa relativamente aos objetivos da pesquisa. A representatividade encontra, em geral, obstáculos no acesso aos textos, na coleta e na uniformidade das regras para a compilação. Diante disto, Viana (2011), apresenta critérios importantes, desenvolvidos a partir das pesquisas embasadas em *corpus*, que buscam garantir, em maior medida, a representatividade:

> Um [*dos critérios*] corresponde à noção de diversidade: um *corpus* que objetiva representar a totalidade de uma língua

> precisa abarcar uma ampla gama de gêneros discursivos, contexos de produção, participantes (de diversas faixas etárias, orignes geográficas, sexos, classes sociais etc.), entre outros. Ao mesmo tempo, a diversidade deve ser temperada com a concepção de equilíbrio (VIANA, 2011, p. 28).

Partindo destes pressupostos da Linguística de *Corpus*, o *corpus* analisado nesta pesquisa teve como base para a compilação o *corpus* CALIBRA (Catálogo da Língua Brasileira), que se pauta pelas questões da busca de representatividade apontadas por Viana (2011).

O CALIBRA é o resultado de uma iniciativa conjunta do Grupo de Pesquisa Produção de Significado em Ambientes Multilíngues da Universidade Federal de Ouro Preto e do Laboratório Experimental de Tradução da Universidade Federal de Minas Gerais. Atualmente, o CALIBRA conta com cerca de 1 milhão de palavras (*tokens*), compilados com base na tipologia do contexto de cultura (HALLIDAY, 1978, p. 139). Esta tipologia é definida segundo cinco variáveis, as quais refletem a organização do modelo metafuncional da língua. São elas:

Especialização: segundo a qual leva-se em conta o fato de o texto ser produzido pelo conhecimento técnico de uma determinada área (especializado/não-especializado).

Papel da língua na situação: segundo o qual o texto é visto como fundamental para a situação sócio-cultural, ou apenas funciona para facilitar a ocorrência da situação (constitutivo/auxiliar).

Modo de produção: que observa a forma pela qual o texto foi originalmente produzido pelo falante (escrito/falado).

Modo de interação: apesar de todo texto ser, a rigor, diálogo, considera o tipo de relação estabelecida textualmente entre os interlocutores, pressupondo a resposta ou não do ouvinte (monólogo/diálogo).

Processo sócio-semiótico: a forma pela qual os textos estão dispersos no contexto de cultura, dividindo-se em explorar, explicar, reportar, recriar, compartilhar, fazer, recomendar, capacitar (Matthiessen *et al*., 2008, p. 190-198).

Os processos sócio-semióticos são, por sua vez, assim caracterizados: o processo explicar envolve o uso da língua como forma de transmissão de conhecimento, que pode se dar tanto entre

pares, quanto do especialista para o leigo. Exemplos de textos pertencentes a este processo são livros didáticos, aulas e palestras. O processo reportar implica no uso da língua como forma de construir linguisticamente um evento que aconteceu no mundo. Como exemplos tem-se reportagens de jornal, biografias e relatórios. O processo recriar busca criar linguisticamente um evento que aconteceu no mundo que, anteriormente, foi codificado por outro processo sócio-semiótico, de forma ficcional. Citam-se como exemplo romances, histórias em quadrinho e causos.

A função principal do processo compartilhar é o estreitamento dos laços sociais. Neste processo sócio-semiótico, tem-se a apresentação e negociação de valores, posicionamentos e ideias com o objetivo de pôr à prova a "geografia social", isto é, a proximidade e a distância entre os membros de uma comunidade. Como exemplos podem ser citados o bate-papo, a fofoca e o blog (diário). No processo fazer, a língua não tem papel principal, mas sim o papel de facilitar a execução de uma atividade não-linguística. Por exemplo, cooperações, procedimentos e instruções. Os tipos de texto associados ao processo recomendar possuem a função de controlar o comportamento dos falantes por meio da língua. Neste tipo de texto encontram-se, por exemplo, comerciais, leis (deveres) e regulamentações.

O processo habilitar procura, por meio da língua, facilitar o comportamento dos falantes em determinada situação. Por exemplo, os folhetos turísticos, as leis (direitos) e o aconselhamento. Por fim, cabe aos tipos de texto relativos ao processo explorar a criação de novos significados que deverão ser postos à negociação com outros membros da comunidade. Por exemplo, os artigos acadêmicos, os editoriais e os estudos críticos.

Na pesquisa apresentada neste trabalho, foram utilizados cerca de 10% do CALIBRA, perfazendo cerca de 100.000 itens (*tokens*), selecionados aleatoriamente dentro de cada variável de representatividade.

5 Etiquetagem dos segmentos linguísticos e extração dos dados

Após a compilação, o *corpus* foi etiquetado visando a identificação de padrões linguísticos. Esta busca de padrões foi feita aliando-se duas perspectivas complementares. A primeira, advinda da Linguística de *Corpus*. A segunda, da teoria linguística base da análise – a teoria sistêmico-funcional.

a) Do ponto de vista da Linguística de *Corpus*, a busca de padrões linguísticos visa sempre à probabilidade de uso. Isto implica na investigação de combinações significativas relativamente ao *corpus*. Viana (2011) afirma:

> Os linguístas de *corpus* não concebem o uso de uma língua como um sistema de possibilidades. Em outras palavras, o objetivo de suas pesquisas não é a descrição das combinações possíveis de serem realizadas (...). Isso signfica que à Linguística de *Corpus* interessam as combinações ou padrões mais prováveis de ocorrerem. É isso, de forma resumida, que corresponde à compreensão de linguagem como probabilidade (VIANA, 2011, p. 39).

Nesta pesquisa, especificamente, o que se buscou foram as formas gramaticais que realizam as funções do Tema Predicado em português brasileiro, 'SER...QUE', 'QUE', 'É QUE', e as combinações significativas, ou quais elementos que com elas co-ocorrem. Objetivou-se, assim, determinar a ocorrência ou não-ocorrência dos elementos no Tema Predicado e, a partir de então, quando ocorrem, a probabilidade de estes itens serem predicados. Buscou-se, igualmente, a distribuição das frequências do Tema Predicado ao longo da tipologia mediante a qual o *corpus* foi compilado (linguagem como probabilidade).

b) Do ponto de vista da teoria sistêmico-funcional, as construções do Tema Predicado foram analisadas dentro da visão trinocular, levando-se em consideração a sua função – oracional e no fluxo discursivo. Para tanto, tomou-se, num primeiro momento o conjunto dos dados extraídos da busca no *corpus* desta pesquisa, comparados às descrições de pesquisas anteriores (HALLIDAY, 2005; MATTHIESSEN, 1995; BRAGA, 1991; LONGHIN e ILARI, 2000) tanto para a identificação do Tema Predicado – pela realização estrutural – quanto para a sua função habilitadora das funções ideacionais e interpessoais.

Para se chegar a resultados que efetivamente contemplassem a complementaridade destes dois pontos de vista, cumpriram-se as seguintes etapas: 1) busca automática das realizações estruturais das funções gramaticais de interesse da pesquisa; 2) análise semi-automática das linhas de concordância resultantes da busca automática para as configurações gramaticais; 3) análise sob a visão trinocular.

A primeira etapa teve início com uma busca automática pelas realizações estruturais das funções gramaticais que codificam os significados de focalização, na forma do Tema Predicado: SER... QUE, QUE, É QUE. Isto foi feito com o auxílio da ferramenta Concord do *software* WordSmith Tools (SCOTT, 2007). Foi, então, elaborada uma lista de buscas do Concord que contemplasse todas as construções dos significados de focalização, mas, complementarmente, de outras construções possíveis de realizá-los gramaticalmente.

A análise das linhas de concordância foi capaz de revelar que um dos elementos característicos que é comum a todas as construções é a presença do verbo copular 'ser'. Com isto, a lista de busca do Concord foi composta por todas as formas finitas e infinitivas de 'ser'. Mediante a busca pelas linhas de concordância entre a lista de buscas e suas respectivas ocorrências no *corpus* da pesquisa, obteve-se uma primeira lista com 16.349 linhas de concordância.

A seguir se apresenta o Quadro 1, que traz uma lista com as dez primeiras linhas de concordância obtidas a partir da busca do Concord, que poderiam ser a realização da função de Tema Predicado:

QUADRO 1
Linhas de concordância para 'SER...QUE', 'QUE', 'É QUE'
no *corpus* da pesquisa

N	Concordance
1	a diminuição desse índice de dosagem alcoólica que... **foi que**... possibilitou ou que influenciou nessa diminuição nos acidentes de trânsito.
2	a-se o narrador e fala, exatamente disso— A: [é. . . **é esse texto que** a gente está estudando— J: Bom, então tá vendo?
3	um dado de elaboração. **era esse que** era o tipo penal.
5	C: Pra esse homem ser assim ele herdou o X agazinho da mãe né? G: Hm.
	C: Então...tá... tem que ser mulher **é só mulher que** interessa.
6	de 1° de maio não é uma unidade de discurso, mas **é uma unidade que** vai se frutificar na luta do povo brasileiro no próximo período
7	ida externa em investimentos educacionais. **É com muita alegria que** nos reunimos nesta noite para festejar a tão sonhada conquista de nossa formatura.
8	atrito de grandes proporções. **É o que** estamos observando hoje.
9	Então **são todas frases correlatas que** me fizeram lembrar uma da outra.
10	meus colegas são aqueles que pensam que 'A Internet é um ícone azul que fica ali no cantinho' (risos). Então, é um pouco difícil, às vezes, explicar o que **é que** é software livre.

Mediante a leitura das linhas, foram eliminadas aquelas não-relacionadas – isto é, as ocorrências de 'ser' nos ambientes que não possuem a forma e/ou significado de focalização; como por exemplo, "*O FIES é um programa do Ministério da Educação*" e "*isso era... é, é... era um golpe de marketing*". Restaram, assim, 321 linhas de concordância cujos segmentos linguísticos possuem a estrutura da realização das funções gramaticais e/ou o significado de focalização, alvos desta pesquisa.

Estas linhas passaram por nova análise e outros 118 segmentos foram descartados por, apesar de possuírem a estrutura de realização das funções de focalização, não formam o padrão linguístico característico deste tipo de função; como por exemplo, "*ele ajudou a gente a subir aquele barranco, ah mas* ***foi*** *terrível essa noite*" e "*a mãe do menino* ***é*** *essa* ***que*** *é separada do rapaz*".

Restaram, por fim, 203 segmentos, os quais se tornaram o efetivo objeto de análise da pesquisa. Estes segmentos foram então anotados de forma semiautomática com o auxílio do *software* UAMTools (O'DONNELL, 2008), com o objetivo de buscar padrões que revelam (a) a distinção entre as diferentes configurações e (b) a caracterização de cada função específica. Estes, então, foram analisados segundo a visão trinocular.

6 Resultados preliminares: a frequência do Tema Predicado em português brasileiro

Como foi mencionado anteriormente, esta se trata de uma pesquisa em andamento. Por este motivo, os resultados são parciais e, neste estágio, são capazes de oferecer apenas dados relativos à abordagem "de baixo", procurando apresentar os diferentes tipos de realização para as funções do Tema Predicado. Para tanto, são apontadas as diferentes estruturas possíveis para esta função, bem como a frequência relativa de cada uma destas estruturas. Somando-se a estes dados, apresenta-se, também, a frequência do Tema Predicado em português brasileiro.

Tendo início com o olhar "de baixo", os resultados obtidos a partir da análise do *corpus* desta pesquisa puderam revelar as ocorrências das realizações na estrutura oracional da gramaticalização da focalização, as quais se distribuem nas funções de predicação como se apresentam na Tabela 1.

TABELA 1
Ocorrências das estruturas que realizam o Tema Predicado no *corpus*

Função:	*Número de ocorrências*	*Número total de sentenças do corpus*	*Frequência relativa quanto ao número total de sentenças para a língua*
Tema Predicado	203	8.787	2,31%

Como foi dito anteriormente, foram utilizados aproximadamente 10% do CALIBRA para a condução deste trabalho; mais precisamente 97.429 itens, distribuídos por 8.787 sentenças, com a razão de 11,08. Da Tabela 1 é possível depreender que, da frequência relativa para a língua, das 8.787 sentenças do *corpus*, 2,31% gramaticalizam a focalização no Tema; o que significa, em média, uma ocorrência de Tema Predicado a cada 202 sentenças.

Quanto à distribuição das diferentes formas de realização do Tema Predicado, os resultados obtidos pela análise do *corpus* desta pesquisa mostram que, estão representadas entre as 203 ocorrências as construções de gramaticalização relativas às construções SER...QUE, QUE, e É QUE, identificadas nas pesquisas anteriores (em particular, Braga, 2009), como por exemplo:

Exemplo 5 - *Isso **foi** a mãe da Ana **que** me contou.*

Exemplo 6 - *Demora para imaginar quando **é que** vai acontecer.*

Exemplo 7 - *Aí as meninas **que** foram lá ver.*

A distribuição das formas de gramaticalização das funções de predicação podem ser vistas na Tabela 2, apresentada a seguir:

TABELA 2
Distribuição das diferentes estruturas do Tema Predicado no *corpus*

Função relativa à realização:	*Número de ocorrências*	*Frequência relativa quanto ao número de ocorrências*	*Frequência relativa quanto ao número total de sentenças para a língua*
SER...QUE	131	64,53%	1,49%
É QUE	43	21,18%	0,49%
QUE	29	14,29%	0,33%
Total	203	100%	2,31%

7 Conclusões preliminares

Este trabalho teve como ponto de partida o ponto de contato entre a Linguística de Córpus e uma teoria linguística descritiva – a teoria sistêmico-funcional. Tendo como objetivo abordar a investigação de uma função gramatical a partir do *corpus*, o Tema Predicado, este trabalho procurou mostrar quais os passos necessários para que se possa extrair dados do *corpus* que sejam relevantes para uma pesquisa de descrição linguística que utilize categorias de uma teoria gramatical consolidada. Os resultados prelçiminares mostram como a investigação dos padrões linguísticos no *corpus* pode contribuir, corroborando os dados encontrados na descrição linguística. Contudo, de forma mais importante, pôde apontar uma contribuição de maior importância da abordagem de *corpus*: o efetivo uso e ambiente funcional da função gramatical, que é, em última instância, a sua distribuição estrutural e frequência ao longo do *corpus*.

Referências

BRAGA, M. L. As Sentenças Clivadas no Português Falado no Rio de Janeiro. *Organon*, Porto Alegre, v. 5, n. 5, p. 109-125, 1991.

BRAGA, M. L. Construções clivadas no português do brasil sob uma abordagem funcionalista. *Matraga*, Rio de Janeiro, v. 16, n. 24, p. 173-196, jan./jun. 2009.

CAFFAREL, A. *A systemic functional approach to grammar of French*: From grammar to discourse. London and New York: Continuum, 2006.

CAFFAREL, A.; MARTIN, J.; MATTHIESSEN, C. (Ed.). *Language typology*: a functional perspective. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2004.

CAGLIARI, L. C. Elementos de fonética do português brasileiro. (1981). Tese (Livre Docência em Linguística) - Departamento de Linguística, Instituto de Estudos Linguísticos, Universidade Estadual de Campinas, Campinas, 2011.

FIGUEREDO, G. *Introdução ao perfil metafuncional do português brasileiro*: contribuições para os estudos multilíngues. (2011). 385 p. Tese (Programa de Pós-Graduação em Estudos Linguísticos) - Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

HALLIDAY, M. A. K. *Language as social semiotic*: the social interpretation of language and meaning. London & Baltimore: Edward Arnold & University Park Press, 1978.

HALLIDAY, M. A. K. *On grammar*. London: Continuum, 2002. (The collected works of M. A. K. Halliday, v. 1).

HALLIDAY, M. A. K. *Studies in English language*. London: Continuum, 2005. (The collected works of M. A. K. Halliday, v. 7).

HALLIDAY, M. A. K.; GREAVES, W. *Intonation in the grammar of English*. London and Oakville: Equinox, 2008.

HALLIDAY, M. A. K.; MATTHIESSEN, C. *An introduction to functional grammar*. 3. ed., London: Edward Arnold, 2004.

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

HERKE-COUCHMAN, M. *SFL, corpus and the consumer*: an exploration of theoretical and technological potential. (2006). 329 p. Tese – Division of Linguistics and Psychology, Department of Linguistics, Macquarie University, Sydney, 2006.

LONGHIN, S.; ILARI, R. Uma leitura hallldayiana das sentenças clivadas do português. *Alpha*, São Paulo, n. 44, p. 193-213, 2000.

MARTIN, J.; ROSE, D. *Working whith discourse*: meaning beyond the clause. 2. ed., London: Continuum, 2007.

MATTHIESSEN, C. *Lexicogrammatical cartography*: English systems. Tokyo: International Language Science Publishers, 1995.

O'DONNELL, M. The UAM CorpusTool: software for corpus annotation and exploration. In: CALLEJAS, Bretones *et al.* (Ed.). *Applied Linguistics Now*: understanding language and mind / la linguística aplicada hoy: comprendiendo el lenguaje y la mente. Almería: Universidad de Almería. 2008. p. 1433-1447.

SCOTT, M. *WordSmith Tools*. Oxford: Oxford University Press, 2007.

VIANA, V. Linguística de *Corpus*. In: VIANA, V.; TAGNIN, S. (Org.). *Corpora no ensino de línguas estrangeiras*. São Paulo: HUB Editorial, 2011.

Mapeamento das orações existenciais no português brasileiro

Adriana Silvina Pagano[1]
Giacomo Patrocinio Figueredo[2]
Kicila Ferreguetti[3]

RESUMO: Este trabalho apresenta um estudo em andamento das orações existenciais no português brasileiro, realizado com subsídios da linguística de corpus. Os dados são gerados através de buscas no corpus CALIBRA (Catálogo da Língua Brasileira), desenvolvido no ICHS/UFOP e na FALE/UFMG, composto por textos representativos de oito processos sócio-semióticos, cada um deles perfazendo 125.000 palavras (tokens), totalizando um milhão de tokens. O desenho do corpus está pautado pelos princípios descritivos da linguística sistêmico funcional, em particular da tipologia de registros baseada no contexto. O objetivo é calcular a frequência de ocorrência de orações existenciais por processo sócio-semiótico e analisar seu papel na desenvolvimento dos textos vinculados a cada processo. O trabalho está inserido nos Grupos de Pesquisa Estudos Multilíngues (UFOP) e Modelagem Sistêmico-Funcional da Tradução e da Produção Textual Multilíngue (Laboratório Experimental de Tradução (LETRA) da FALE/UFMG).

PALAVRAS-CHAVE: orações existenciais, corpus monolíngue, português brasileiro, descrição linguística, linguística sistêmico-funcional.

[1] Doutora em Letras. Professora Associada da Universidade Federal de Minas Gerais. apagano@ufmg.br.

[2] Doutor em Linguística Aplicada. Professor Adjunto de Linguística Aplicada. Universidade Federal de Ouro Preto. giacomojakob@yahoo.ca.

[3] Mestranda em Estudos Linguísticos (Estudos da Tradução). Universidade Federal de Minas Gerais. kicilaferreguetti@yahoo.com.br.

ABSTRACT: This paper reports on an ongoing study of existential clauses in Brazilian Portuguese based on data retrieved from a monolingual corpus designed by researchers at the Federal University of Ouro Preto (UFOP) and the Federal University of Minas Gerais (UFMG). The corpus CALIBRA (Catálogo da Língua Brasileira) is made up by texts deemed representative of eight socio-semiotic processes. Each process counts 125,000 words (tokens), all of them totalling one million tokens. The corpus design draws on descriptive principles within systemic functional linguistics, more precisely on a context-based register typology of texts. The aim is to compute the frequency of occurrence of existential clauses in each type of socio-semiotic process and analyse the role played by existentials in the texts belonging to each process. The study is developed as part of the activities of the research groups Multilingual Studies (UFOP) and Systemic-Functional Modelling of Translation and Multilingual Text Production (Laboratory for Experimentation in Translation – LETRA, at UFMG).

KEYWORDS: existential clauses, monolingual corpus, Brazilian Portuguese, language description, systemic-functional linguistics.

1 Introdução

Na linguística sistêmico-funcional (HALLIDAY; MATTHIESSEN, 1999; 2004), as orações existenciais constroem significados relativos ao surgimento ou aparecimento de seres ou entidades no discurso e ao acontecimento de processos e eventos. Uma de suas principais funções é a de introduzir um novo referente no discurso, o qual será retomado e sobre o qual outros significados serão construídos, uma vez que o mesmo passa a existir. Para Halliday e Matthiessen (2004), as orações existenciais cumprem um papel relevante em determinados tipos de texto, como é o caso das histórias, nas quais introduzem participantes no movimento inicial da narrativa, ou em guias turísticos, nos quais apresentam lugares e pontos de interesse sugeridos para visitação do público leitor. Além desses dois tipos de texto, presume-se que os significados existenciais desempenhem papeis relevantes em outros tipos textuais, embora estudos a respeito sejam praticamente inexistente na literatura. No que diz respeito ao repertório textual dos usuários da língua portuguesa do Brasil, pouco se conhece sobre a ocorrência e características das orações existenciais em diferentes tipos textuais. Este trabalho visa preencher essa lacuna através de um estudo exploratório das orações existenciais no português brasileiro, utilizando-se subsídios da linguística de corpus.

Os dados são gerados através de buscas no corpus CALIBRA (Catálogo da Língua Brasileira), desenvolvido desenvolvido no ICHS/UFOP e na FALE/UFMG, composto por textos representativos de oito processos sócio-semióticos, cada um representado no corpus por textos que perfazem 125.000 palavras (tokens), totalizando um milhão de tokens. O objetivo é mapear a frequência de ocorrência de orações existenciais por processo sócio-semiótico e analisar seu papel na desenvolvimento dos textos vinculados a cada processo. O trabalho está inserido no Grupo de Pesquisa em Estudos Multilíngues do departamento de Letras ICHS/UFOP e e no grupo de pesquisa Modelagem sistêmico-funcional da tradução e da produção textual multilíngue, desenvolvido no Laboratório Experimental de Tradução (LETRA) da Faculdade de Letras/UFMG.

2 Revisão teórica

No marco teórico da linguística sistêmico-funcional (MATTHIESSEN, 1995; HALLIDAY; MATTHIESSEN, 2004), as orações existenciais são definidas como aquelas que constroem uma entidade ou evento que é apresentado como Existente. O Existente pode ser postulado em relação a um espaço (concreto ou simbólico) ou a um acontecimento. Nesse sentido, as orações existenciais geralmente possuem uma circunstância de localização temporal ou espacial. Os exemplos, a seguir, extraídos do Collins WordBanks Online, ilustram as características apontadas em relação às orações existenciais da língua inglesa. Os verbos lexicais e estruturas realizadoras de existência estão sublinhadas e o Existente está grifado em negrito:

> There was **an uncomfortable pause**.
>
> There was **a morgue assistant** there, a young guy with tied-back hair and a goatee beard, perhaps a student.
>
> A door opened. There was **the sound of laughter**.
>
> There are **regular, direct services to London**, with the journey taking approximately one hour.
>
> **Fertility clinic malpractice** exists too
>
> Yes, **performance enhancing drug-taking** exists in the sporting world, in cycling and consequently in the Tour.[4]

[4] Disponível em: < http://wordbanks.harpercollins.co.uk/>. Acesso em: 17 maio 2012.

Como pode ser visto nos exemplos acima, em inglês, as orações existenciais são tipicamente realizadas pela construção THERE + verbo TO BE e pelo verbo EXIST. Todavia, outros verbos lexicais também realizam processos existenciais, os quais são agrupados de acordo com os seguintes tipos (HALLIDAY; MATTHIESSEN, 2004: 258):

QUADRO 1
Verbos que realizam Processos em orações existenciais
Fonte: Halliday & Matthiessen, 2004: 258. Nossa tradução.

Tipos		Verbos
neutro	existência acontecimento	*exist, remain* *arise, occur, come about, happen, take place*
com significado circunstancial	tempo lugar	*follow, ensue* *sit, stand, lie, hang, rise, stretch, emerge, grow*
abstrato		*erupt, flourish, prevail*

É oportuno destacar que, na língua inglesa, o item THERE cumpre a função de Sujeito da oração, função esta obrigatória no sistema interpessoal dessa língua. Não possui, como Halliday e Matthiessen (2004) assinalam, função na transitividade da oração. Seu papel é o de indicar que será apresentado um Existente. Do ponto de vista da construção de significados textuais, por meio da construção com THERE, o processo existencial ocupa posição temática na oração e seu papel é apresentar um novo referente no discurso (MATTHIESSEN, 1995).

Uma característica adicional das orações existenciais na língua inglesa é que mesmo quando realizadas por outros verbos lexicais além de "to be", podem ser realizadas pela estrutura THERE + verbo lexical encenando o processo existencial. Nesses casos, o item THERE, ocupando posição temática, contribui para indicar que será apresentado um Existente. Os exemplos, a seguir, extraídos do Collins WordBanks Online, ilustram esse aspecto. Os verbos lexicais e estruturas realizadoras de existência estão sublinhadas e o Existente está grifado em negrito. As construções com THERE, como o contraste evidencia, permitem colocar o Existente em posição remática:

Sameth felt the smile spread across his face as **applause** erupted in the stands.

More immediately, however, in mid-August 1969, there erupted **the political volcano known as the Matesa scandal**.

Here, **death and life** flourished side by side

Little did we realise that, beyond the public face of Italian and Greek-owned cafes ubiquitous in the 1960s, there flourished **a culinary sensibility that would transform the food Australians ate**.

When he reached home, Joanna had been furious. It had probably been out of relief that he was safe, but **a quarrel** ensued.

There ensued **an animated conversation on deck, part in German, the balance of it in French**.

Such polar pairings, while **they** do not happen in nature, nevertheless represent a significant symbolic balancing within the ecological order itself.

Yes, it freezes over usually for several months, up to five to six months, so there happens **next to nothing** during the Winter time, except for bacteria activity which releases all the nutrients so that the table is turned when the Spring comes.

Excitement grew in her as she realized this was the beginning of the magic.

And even beyond that, there grew **the concern that if the lake was allowed to continue to deteriorate and eventually die, it would seriously damage the self-esteem of the city**.[5]

Na língua inglesa, o Existente é geralmente um ente não específico, característica que diferencia, por exemplo, orações existenciais realizadas pelo verbo lexical "to be" de orações relacionais circunstanciais, como os exemplos a seguir mostram. O Existente, ente não específico, está grifado em negrito:

On the wall is **a picture**, isn't there – Oração existencial

On the wall is the picture, isn't it – Oração relacional circunstancial

[5] Disponível em: < http://wordbanks.harpercollins.co.uk/>. Acesso em: 17 maio 2012.

Todavia, o Existente pode ser também um ente específico, como demostra Davidse (1992). Isso se deve à dupla função do grupo nominal no sistema linguístico. Por um lado, expressa a classificação de um dado ente e, por outro, ancora esse ente no discurso desenvolvido no texto, apresentando-o como conhecido ou desconhecido para o interlocutor. Assim, Davidse (1999) identifica orações existenciais cardinais e orações existenciais enumerativas e caracteriza o tipo ente que pode ser Existente desses dois tipos de orações existenciais.

As orações existenciais cardinais indicam o caráter cardinal de uma dada instanciação ou de um tipo expresso pela especificação do Existente. Possuem apenas um componente obrigatório, um grupo nominal que realiza o Existente. O item THERE pode ser utilizado, mas não é um item obrigatório. Também não é obrigatória a especificação de uma circunstância de localização espacial ou temporal. O Existente é um geralmente um ente não específico. Os exemplos a seguir ilustram orações existenciais cardinais:

> **Storms of protest** followed.
>
> Throughout the borderlands – Poland, Baltic, Finland and the Caucasus – there followed **widespread unrest and insurrection**.[6]

As orações existenciais enumerativas encenam uma enumeração de instâncias que se relacionam a um hiperônimo passível de ser construído co- e contextualmente. O Existente é geralmente um ente específico, o qual pode ser tanto genérico como definido. Os exemplos a seguir ilustram orações existenciais enumerativas:

> There's **humor, giddiness, absurdity, anger** ... it's all mixed in and it's never when you think it would be.
>
> Roy only drank mate, a form of health beverage – particularly disgusting – alleged to promote longevity. Frances usually drank coffee, Colombian or Blue Mountain; then there were **the teas**, Lapsang for Lady Kathleen, in case she should call (she never did), Darjeeling for Roy 's boyfriend, Typhoo for Painter. Luckily there was, in addition, **the range of health-promoting herbal teas left by the receptionist before last**, whose boyfriend had been – very likely still was – a drug pedlar.[7]

[6] Disponível em: < http://wordbanks.harpercollins.co.uk/>. Acesso em: 17 maio 2012.

[7] Disponível em: < http://wordbanks.harpercollins.co.uk/>. Acesso em: 17 maio 2012.

Os grupos nominais que realizam os Existentes nos exemplos acima se relacionam a um hiperônimo que pode ser recuperado co- e contextualmente. No primeiro caso, trata-se de uma enumeração de sentimentos; já no segundo, os Existentes podem ser relacionados ao hiperônimo "bebidas".

Diferentemente das orações existenciais cardinais, as orações existenciais enumerativas em inglês precisam ser obrigatoriamente realizadas pela construção com o Sujeito funcional THERE.

Um dado adicional apresentado pelos teóricos da linguística sistêmico funcional sobre as orações existenciais na língua inglesa é sua baixa frequência de ocorrência, em comparação com outros tipos de orações (MATTHIESSEN, 1995).

Estudos sobre as orações existenciais no português brasileiro possuem ainda caráter incipiente. Dentre eles, destaca-se o estudo de Franchi, Negrão e Viotti (1998), no qual os autores abordam as orações existenciais do ponto de vista da semântica discursiva. Embora o referencial teórico utilizado pelos autores seja diferente daquele aqui adotado, apontamentos feitos pelos mesmos oferecem dados que nos permitem construir um percurso metodológico inicial desta pesquisa. Dentre eles, temos uma relação de verbos lexicais que realizam significados existenciais no português brasileiro. Os autores apontam como verbos prototípicos "ter", "haver" e "existir", seguidos por uma classe de verbos caracterizados por operarem numa estrutura na qual o Sujeito está posposto ao verbo. Os autores citam como exemplos os verbos "acontecer", "aparecer", "chegar", "faltar", "ir", "ocorrer", "sobrar", "surgir" e "vir".

Como a seção de Metodologia mostra, os apontamentos de Franchi, Negrão e Viotti (1998) foram levados em conta para a realização de buscas iniciais no corpus CALIBRA.

3 Metodologia

O presente estudo utilizou dados do Corpus CALIBRA – Catálogo da Língua Brasileira, desenvolvido conjuntamente por pesquisadores da FALE/UFMG e do ICHS/UFOP. Trata-se de um corpus monolíngue de português brasileiro falado e escrito, compilado com base numa tipologia de textos pautada por registros sob a perspectiva do contexto de cultura (MATTHIESSEN; TERUYA; WU, 2008). Os textos são classificados de acordo com cinco variáveis:

- Grau de especialização da linguagem (especializada/não especializada)
- Papel da linguagem na interação (ancilar/constitutivo)
- Modo de produção (escrito/falado)
- Modo de interação (monólogo/diálogo)
- Processo sócio-semiótico (explorar; explicar; relatar; recriar; compartilhar; fazer; recomendar; capacitar)

Para a realização deste estudo preliminar, de caráter prospectivo, foram adotados os seguintes procedimentos metodológicos.

Primeiramente, foi elaborada uma lista lemas correspondentes a verbos lexicais passíveis de realizar significados existenciais no português brasileiro. Como dito acima, os apontamentos de Franchi, Negrão e Viotti (1998) fornecerem alguns subsídios para a elaboração de tal lista, a qual foi complementada com verbos sinónimos e correlatos. O Gráfico 2 a seguir mostra as expressões de busca utilizadas.

QUADRO 2
Expressões de busca utilizadas para a extração de linhas de concordância

acab*	acontec*	aconteç*	adv*	aparec*	apont*	brot*	chegu*
comec*	comparec*	dá	dão	dará	darão	daria	dariam
decorr*	deram	desabroch*	desaparec*	despont*	deste	deu	deu
dur*	durar	é	eclod*	encontr*	era	eram	esvanec*
evanec*	evapor*	exist*	expir*	extingu*	falt*	fic*	foi
fora	foram	foram*	forem	fosse*	fôsse*	há	haja
hajam	hão	have*	haver	havi*	haví*houve*	houvé*	ia
iam	inici*	ir	irá	irão	jaz*	manifest*	morr*
nasc*	perdur*	permanec*	persever*	persist*	pint*	prevalec*	rai*
rebent*	resist*	rest*	rol*	romp*	sai	são	segu*
ser	será*	serão	serem	seria	seriam	sobr*	sum*
surg*	tem	têm	tenha	tenham	ter	terá	terão
teria	teriam	tinha	tinham	teve	tiveram	vai	vão
vir	vem	vêm	vinha	vinham			

Com base nas expressões no Quadro 2, foram extraídas linhas de concordância com o software Wordsmith Tools (SCOTT, 2007). As linhas obtidas foram analisadas manualmente de forma a verificar se as mesmas não correspondiam a ocorrências de outros tipos de orações que não configurassem significados existenciais. Isto foi necessário sobretudo para os lemas dos verbos "ser", "ter" e "haver", uma vez que os dois primeiros possuem uma frequência alta de ocorrência em orações relacionais de identidade, atribuição e posse, e o último em grupos verbais que realizam tempos verbais no passado.

As linhas de concordância obtidas depois de terem sido descartadas aquelas que não representavam significados existenciais foram contabilizadas de forma a se obter uma proporção de sua ocorrência nos tipos de processo sócio-semióticos examinados, a saber, capacitar, compartilhar e fazer. Posteriormente, as ocorrências foram agrupadas inicialmente em padrões visíveis de semelhança estrutural. A seguir, são apresentados resultados dessa análise preliminar.

4 Resultados preliminares

A Tabela 1 a seguir apresenta os dados quantitativos da amostra do corpus CALIBRA analisada.

TABELA 1
Dados quantitativos da amostra analisada

Processo Sócio-Semiótico	Número de textos	Número de linhas concordância obtidas (sem duplicados)	Número de linhas de concordância com significados existências selecionadas manualmente
CAPACITAR	71	2836	174
COMPARTILHAR	61	3539	276
FAZER	62	1871	139
TOTAL	194	8246	509

Como os dados da Tabela nos permitem calcular, a proporção aproximada de orações existenciais selecionadas manualmente em

relação ao total de linhas de concordância obtidas de forma automática é de 6,17%. Isso aponta para a considerável demanda de análise manual dos resultados obtidos, juntamente para a necessidade de se encontrar parâmetros que permitam filtrar as linhas obtidas.

Nos processos sócio-semióticos examinados , os verbos lexicais que realizam significados existenciais mais frequentes foram aqueles apontados como prototípicos em Franchi, Negrão e Viotti (1998), isto é, "haver", "ter" e "existir".

Em relação as características das orações existenciais no português brasileiro, foi possível se chegar às seguintes constatações, a seguir acompanhadas de exemplos retirados da amostra analisada.

As orações existenciais constroem o significado de que algo existe ou acontece tendo um único Participante, o Existente. São tipicamente realizadas por "haver", "ter" e "existir" e, em menor escala, por "ser" e outros verbos.

> É pode ter sido responsável por isso tudo, claro que sim, não tem **a menor dúvida**. Isso é sempre o risco de todo mundo que toma anabolizante, que toma esse tipo d...

Do ponto de vista da construção de significados textuais, o processo existencial ocupa posição temática na oração – papel de apresentação de um novo referente no discurso (Matthiessen, 1995)

> ... relevante a participação do aeroporto de Salvador no desenvolvimento turístico da capital baiana e de todo o Estado da Bahia. São **80 vôos regulares diários para todas as Capitais do País, principais cidades brasileiras e alguns municípios baianos**, ch...

Circunscrevem a existência a um espaço e tempo – orações com circunstâncias de localização temporal e espacial ou orações não finitas

> ... qual é a maior diferença entre elas? Tem **mais liberdade** na internet?

Alguns dos significados parecem situar-se no limiar entre processos, como é o caso de significados existenciais que podem ser interpretados como materiais

> ... Mas rolou depois **uma conversa por telefone**... Nelito – Não foi bem uma convers...

5 Conclusões e futuros direcionamentos da pesquisa

Como foi apontado ao longo deste trabalho, a análise empreendida possui caráter prospectivo, no sentido de se implementar um percurso metodológico que possa ser profícuo para a análise em maior escala, abrangendo todos os processos sócio-semióticos do corpus CALIBRA. Constituem etapas subsequentes àquelas já realizadas, a anotação das linhas de concordância selecionadas no software UAM Tools (O'DONNELL, 1997) de forma a caracterizar as orações existenciais sob a perspectiva da realização das funções na ordem da oração (Processo e Participante) em relação ao grupo verbal e ao grupo nominal que realizam tais funções. Busca-se observar quais grupos nominais realizam prototipicamente os Existentes e em que medida os apontamentos de Davidse (1992; 1999) podem ser mapeados no português brasileiro, isto é, como são realizados os significados existenciais cardinais e enumerativos na nossa língua.

Uma segunda perspectiva de análise diz respeito à função discursiva dos processos existenciais em cada tipo de processo sócio-semiótico, isto é, em que etapas do desenvolvimento do discurso os significados existenciais tendem a ocorrer.

Por último, uma terceira perspectiva de análise tenciona observar em que medida a ocorrência de significados existenciais pode ser correlacionada aos outros tipos de significado e quais as implicações de se escolher este tipo de construções em termos do tipo de representação que se quer construir do mundo externo e do mundo interno da nossa consciência.

Referências

DAVIDSE, K. Existential constructions: a systemic perspective, *Leuvense Bijdragen* 81, p. 71-99, 1992.

DAVIDSE, K. The semantics of cardinal versus enumerative existential constructions. *Cognitive Linguistics,* v. 10, n.3, p. 203-250, 1999.

FIGUEREDO, G. *Introdução ao perfil metafuncional do português brasileiro*: contribuições para os estudos multilíngues. (2011). 385 p. Tese - Programa de Pós-Graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

FRANCHI, C; NEGRAO, E.; VIOTTI, E. Sobre a gramática das orações impessoais com ter/haver. *DELTA*, São Paulo, v. 14, n. spe, 1998.

HALLIDAY, M. A. K.; MATTHIESSEN, C. *An introduction to functional grammar*. 3. ed., London: Edward Arnold, 2004.

HALLIDAY, M. A. K.; MATTHIESSEN, C. *Construing experience through meaning. A language-based approach to cognition.* London, New York: Cassell, 1999.

MATTHIESSEN, C. *Lexicogrammatical cartography*: English systems. Tokyo: International Language Science Publishers, 1995.

MATTHIESSEN, C.; TERUYA, K.; WU, C. Multilingual studies as a multi-dimensional space of interconnected language studies. In: WEBSTER, J. (Ed.). *Meaning in Context*: implementing intelligent applications of language studies. London and New York: Continuum, 2008.

O'DONNELL, M. The UAM CorpusTool*:* software for corpus annotation and exploration. In: CALLEJAS, Bretones *et al.* (Ed.). In: *Applied Linguistics Now*: understanding language and mind / la linguística aplicada hoy: comprendiendo el lenguaje y la mente. Almería: Universidad de Almería, 2008. p. 1433-1447.

SCOTT, M. *WordSmith Tools*. Oxford: Oxford University Press, 2007.

V TRADUÇÃO, COMPARAÇÃO INTERLINGUÍSTICA, ENSINO DE LÍNGUAS

O vocabulário do horror: uma análise contrastiva bilíngue baseada em *corpus* do léxico especializado da série *Supernatural*

Raphael Marco Oliveira Carneiro[1]

RESUMO: Este texto tem como objetivo apresentar, de modo geral, alguns aspectos teóricos e metodológicos utilizados para a análise terminológica de um *corpus* composto por legendas de uma série de ficção chamada *Sobrenatural*. Após uma introdução sobre o que motivou a realização desta pesquisa, apresentaremos o quadro teórico-metodológico em que este estudo se insere e, a seguir, apresentaremos alguns dos passos necessários para a criação de um vocabulário especializado numa plataforma denominada VoTec (FROMM, 2007).

PALAVRAS-CHAVE: Ficção; Linguística de *Corpus*; Terminografia; Terminologia; Tradução.

ABSTRACT: This paper aims to present some of the theoretical and methodological aspects used to carry out a terminological analysis of a corpus composed of subtitles of a fictional TV series named *Supernatural.* After an introduction about what motivated this research, we intend to present some of the study realms which this research is part of. Following this, there will be some of the steps taken in order to generate a specialized vocabulary in a plataform named VoTec (FROMM, 2007).

KEYWORDS: Corpus Linguistics, Fiction, Terminography, Terminology, Translation.

1 Introdução

Esta pesquisa tem como tema o estudo de terminologia na ficção. Mais especificamente, nosso objeto de estudo está concentrado em uma das mais assistidas séries de TV reconhecida mundialmente:

[1] Graduando em Letras – Habilitação em Inglês e Literaturas de Língua Inglesa no âmbito do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia; orientado pelo Prof. Dr. Guilherme Fromm.
e-mail: raphael.olic@gmail.com.

Supernatural/Sobrenatural. Desse modo, analisaremos como a terminologia dessa série é caracterizada dentro de sua área específica que se enquadra na de ficção fantástica caracterizada pela presença de elementos de histórias de horror.

Além de realizar uma análise terminológica, pretende-se levar em conta o modo como o léxico de *Supernatural* foi traduzido. Para que isso fosse feito, planejou-se um *corpus* composto pelas legendas em inglês e português das seis temporadas de *Supernatural*, sendo, pois, porções de falas transcritas produzidas por falantes nativos e de suas traduções, realizadas por brasileiros, compreendidas em um período de tempo. Temos então uma amostra finita da linguagem, de conteúdo especializado, em que os textos são comparáveis.

Um dos motivos que nos levaram à pesquisa proposta foi o fato de que tradutores são frequentemente desafiados pela presença de terminologias de áreas cada vez mais específicas. Esses, por não estarem familiarizados com determinada área, geralmente encontram dificuldades para traduzirem termos específicos. Por isso, justifica-se o estudo de terminologias, a fim de que, a partir dos termos repertoriados em glossários e bancos de dados, o tradutor possa ter à sua disposição ferramentas que facilitem e melhorem a qualidade de sua prática tradutória.

Isso posto, propomos os seguintes objetivos:

a. Compilar um *corpus* contrastivo bilíngue, a partir das legendas em inglês e português da série *Supernatural*;

b. Repertoriar e analisar o léxico especializado em inglês e português;

c. Elaborar um vocabulário bilíngue a ser disponibilizado por meio do site VoTec: Vocabulário Técnico Online (FROMM, 2007).

A seguir serão contemplados, resumidamente, os campos de estudos nos quais esta pesquisa se fundamenta.

2 Linguística de *corpus*, terminologia, terminografia, tradução e literatura fantástica

Algumas áreas estão diretamente relacionadas ao tema proposto desta pesquisa. São elas Linguística de *Corpus*, Terminologia/ Terminografia e Tradução.

> A Lingüística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador (BERBER SARDINHA, 2004, p. 3).

Percebe-se, assim, o uso central e os benefícios gerados pela tecnologia no desenvolvimento de pesquisas que se utilizam de *corpora*.

Essa tecnologia, utilizada em favor da Linguística de *Corpus*, permite não só o armazenamento de *corpora*, como também a sua exploração. Por isso, as ferramentas computacionais que se disponibilizam à análise de *corpus*, como o *WordSmith Tools*, são fundamentais. Esse programa tem sido uma das ferramentas mais versáteis utilizadas na análise linguística, constituindo um dos mais completos conjuntos de ferramentas para os estudos realizados com *corpora* eletrônicos (BERBER SARDINHA, 2004).

Basicamente, o *WordSmith Tools*, criado em 1996 por Mike Scott, da Universidade de Liverpool, Reino Unido, é um conjunto de ferramentas integradas (*Concord, KeyWords* e *WordList*) utilizadas para análise linguística. Ele permite fazer análises baseadas nas frequências e coocorrências de palavras em *corpora* (BERBER SARDINHA, 2009).

Outra área de estudos linguísticos contemplada nesta pesquisa é a Terminologia, que se preocupa principalmente com o estudo de termos de uma dada linguagem de especialidade. Tais termos surgem da necessidade de se denominar conceitos, objetos, processos de diferentes campos do saber. Esse tipo de produção linguística é verificado no universo das ciências, das técnicas e das várias atividades profissionais (FINATTO; KRIEGER, 2004).

Desse modo, o que definitivamente caracteriza uma linguagem como sendo especializada é o léxico. É a seleção lexical que a distingue da língua comum, e por isso os termos tornam-se sua característica preponderante. Assim, o termo é tematicamente marcado, visto que constitui a unidade lexical da linguagem de especialidade.

Enquanto a Terminologia apresenta uma face mais teórica, voltada para as implicações linguísticas, conceituais e comunicativas que subjazem aos termos, a Terminografia é caracterizada por uma abordagem prática. Isso nos leva ao fato de que a Terminografia é responsável pelo "conjunto de práticas e métodos utilizados na compilação, descrição, gestão e apresentação dos termos de uma

determinada linguagem de especialidade" (ALMEIDA, 2010), ou seja, é o campo que estuda a geração de produtos terminológicos, como glossários, dicionários técnicos e bancos de dados.

A Tradução, por sua vez compreende tanto a interpretação quanto a dublagem e a legendagem, além de investigar questões práticas como treinamento de tradutores assim como critérios de avaliação de traduções (BAKER, 1998 apud OLOHAN, 2004). Nesse sentido, o *corpus* é visto primeiramente como uma ferramenta de pesquisa que possibilita o estudo da tradução de várias maneiras. Pode-se citar como uma dessas maneiras, o uso de *corpora* paralelo para a extração de terminologias.

Bowker e Pearson (2002) também fazem referência ao uso de *corpora* ressaltando suas contribuições para a identificação de termos especializados, bem como para a identificação do co-texto de um termo que pode fornecer definições e descrições do conceito a que se refere.

Para um embasamento na área em que o léxico especializado de *Supernatural* está inserido, um estudo também foi feito sobre a Literatura Fantástica, a partir da obra de Todorov (2008).

3 Considerações sobre a área temática de *Supernatural*

Como um dos passos para a construção de repertórios terminológicos utiliza-se a elaboração de árvores de domínios para que se estabeleça uma relação conceitual e semântica dos termos que constituem dada especialidade.

> O princípio fundamental da *terminologia* é a pertinência dos *termos* a áreas temáticas, estruturadas em sistemas de classificação de conhecimentos especializados. Cada especialidade apresenta um sistema de áreas, denominado também *árvore temática*, que deve aparecer evidente em qualquer fundo de terminologia coerente (NOLET, D.; PAVEL, S., 2002, p. 1).

A partir da citação acima, percebe-se a necessidade de uma árvore de domínio ou temática para a execução do trabalho terminológico. Desse modo, a partir do estudo feito por Todorov (2008) e do sistema de classificação do CNPq das áreas do conhecimento chegamos à seguinte árvore de domínio:

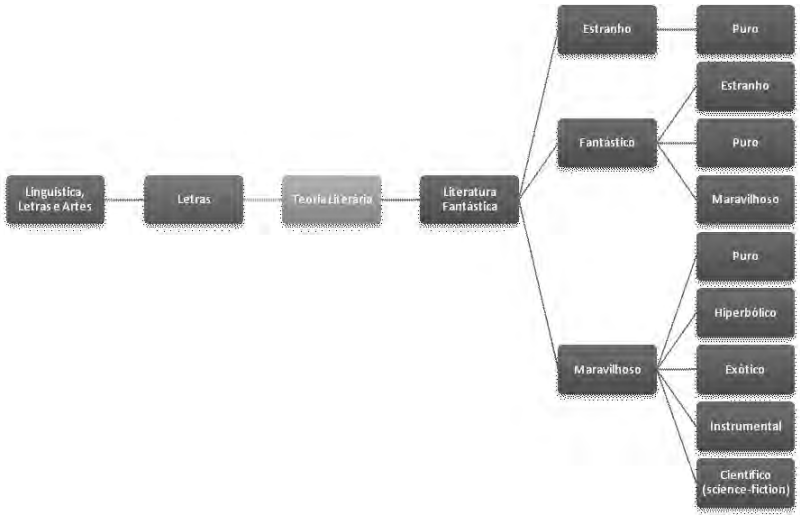


FIGURA 1 - Árvore de domínio da Literatura Fantástica

É importante lembrar que essa árvore de domínio não representa as relações conceituais dos termos da série. Esses domínios representam uma série de características encontradas em histórias de horror de um modo geral. Isto é, a partir deles podemos determinar a área ficcional, na qual *Supernatural* está inserida.

De acordo com os estudos já feitos, podemos dizer que *Supernatural* apresenta características de dois dos domínios apresentados acima. São eles: Maravilhoso Instrumental e Maravilhoso Científico. O Maravilhoso Instrumental refere-se a objetos de origem mágica ou que permitem a comunicação ou contato com o mundo sobrenatural. Consideramos também como parte desse domínio, objetos que são utilizados para o combate às criaturas sobrenaturais, como sal grosso e bala de prata. O Maravilhoso Científico refere-se às histórias em que o sobrenatural é explicado de maneira racional, mas partindo de leis que a ciência desconhece. Isto é, o conhecimento sobrenatural é sistematizado e explicado criando uma racionalidade própria do universo ficcional. Essa característica contribui ainda mais para a organização de um conhecimento terminológico.

Apesar de não se tratar de uma obra literária em si, pelo contrário, trata-se de uma produção audiovisual, a Literatura se interpenetra em *Supernatural.* Muito do que está representado em *Supernatural* tem a sua fonte em contos fantásticos, folclore, contos

de fadas e histórias populares, o que caracteriza a sua face literária. Assim, pode-se afirmar que a série *Supernatural,* levando em conta a intertextualidade e a interface entre as artes, está inserida em uma subárea específica da Literatura, a saber, a Literatura Fantástica. As produções artísticas, sejam elas literárias ou audiovisuais realizadas nesta subárea, por apresentarem conceitos, objetos e características próprias, utilizam um campo lexical muito peculiar, que é caracterizado acima de tudo pelo seu alto caráter abstrato, geralmente relacionado ao sobrenatural. Isto é, os conceitos representados pelos termos estão presentes no imaginário, não há uma realidade concreta que se aplique a eles. Não há referenciais reais dos conceitos veiculados pelos termos em *Supernatural.*

Nota-se, então, que não se trata de uma área técnica-científica. Mesmo não sendo uma área científica, não deixa de ser uma área caracterizada por um conjunto de termos que representam e transmitem um dado conhecimento, ou seja, há uma terminologia. Isso revela que as terminologias não são de uso exclusivo das ciências e técnicas. Por esse mesmo fato é que se torna interessante pensar no caráter sistemático desse conhecimento apresentado cada vez maior dentro da série em análise.

Nesse sentido, o grande desafio seria organizar uma relação entre os conceitos dentro de um universo tão amplo quanto a ficção. A seguir, apresentaremos uma primeira tentativa na organização do vocabulário da série em campos semânticos e lexicais.

4 O vocabulário do horror: conceitos e caracterização

Partindo dos conceitos lexicográficos de vocabulário, campo semântico e campo lexical, pudemos caracterizar o que se convencionou chamar de vocabulário do horror.

Levando em conta que "o vocabulário busca ser representativo de um universo de discurso" (MULLER, 1968 apud BARBOSA, 1995 apud WELKER, 2004, p. 24), o léxico em *Supernatural* é representativo do contexto discursivo em que os personagens estão inseridos, que se caracteriza pela presença de situações relacionadas ao sobrenatural. Percebe-se, então, que as palavras usadas no discurso das personagens revelam o contexto marcado pela atividade deles: caçar seres sobrenaturais.

"Chamamos de *campo semântico* o conjunto de lexias [lexemas] que têm um mesmo componente semântico identificador de campo." Esta é a definição dada por Mel'cuket al.(1995) citada por Welker (2004). Desse modo, "um campo lexical é uma estrutura paradigmática constituída de unidades lexicais que repartem uma zona de significação comum e que se encontram em oposição imediata umas com as outras (COSERIU, 1977 apud FAULSTICH, 1980 apud WELKER, 2004, p. 33).

A partir desses dois conceitos foi possível identificar os campos semânticos e lexicais do vocabulário do horror. Por exemplo, no campo semântico *monstros*, encontramos um campo lexical formado por *Wendigo, metamorfo, lobisomem.* Percebe-se, então, que ao denominador comum, ou componente semântico identificador de campo, muitas vezes corresponde um hiperônimo (arquilexema), por exemplo, *monstros.*

Enfim, o vocabulário do horror é constituído pelos campos semânticos, bem como pelos seus campos lexicais. É importante lembrar que o vocabulário do horror não se restringe aos campos lexicais encontrados nesta pesquisa, visto que as histórias de horror são muito vastas e pesquisas que levem em conta outras obras de horror certamente encontrariam outros campos lexicais.

A seguir, apresentamos alguns campos semânticos com seus respectivos campos lexicais a título de exemplificação:

QUADRO 1
Alguns campos semânticos e lexicais de *Supernatural*

Bruxas	shtriga
Demônios	Daeva, achiri
Deuses	Vanir
Doenças/Vírus	Croatoan
Espíritos	poltergeist, a Woman in White (umaMulher de Branco), tulpa, Bloody Mary (Mary Sangrenta)
Instrumentos	rock salt (sal grosso), silver bullet (bala de prata)
Línguas	Enochian (Enoquiano)
Monstros	shapeshifter (Metamorfo), hellhound (cão do inferno), Wendigo, werewolf (lobisomen), skinwalker (transmorfo), crocotta, djinn, changeling, rougarou, ghoul
Mortos-vivos	vampire (vampiro), zombie (zumbi)
Semi-deuses	trickster

Os termos entre parêntese referem-se aos equivalentes em português encontrados no *corpus*. Os termos que não apresentam equivalentes são aqueles que foram usados como empréstimos na tradução (AUBERT, 1998).

5 Procedimentos de análise

A seguir, apresentaremos o planejamento feito e os critérios tipológicos adotados para a compilação de nosso *corpus* de estudo, de acordo com os estabelecidos por Berber Sardinha (2004).

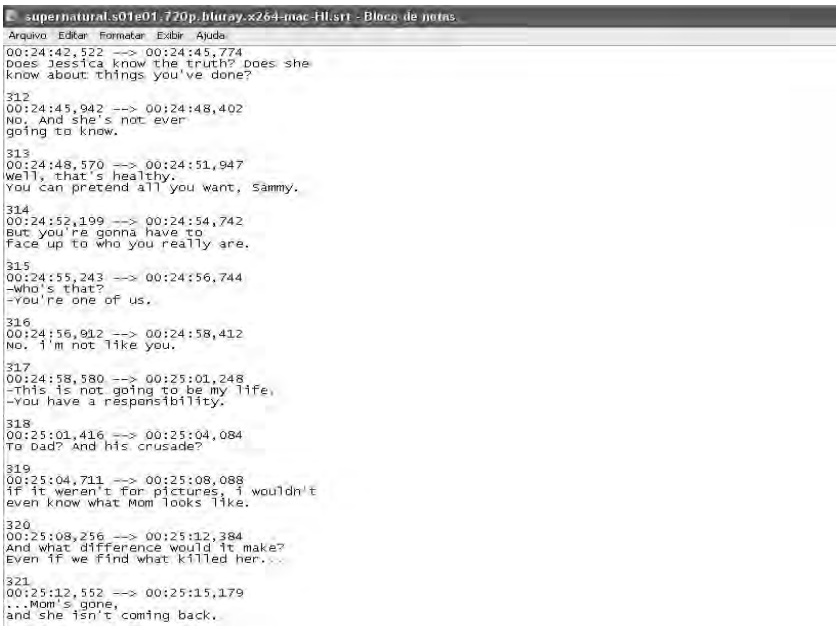
QUADRO 2
Tipologia do *Corpus* de estudo

Modo	Escrito (legendas disponíveis na Internet)
Tempo	Sincrônico (legendas de 2005 - 2011)
Seleção	Amostragem (linguagem de textos da área de ficção)
	Estático (seleção não renovável)
Conteúdo	Especializado (legendas de uma série de ficção)
Autoria	Língua nativa (inglês americano e tradução em português brasileiro)
Disposição Interna	Paralelo (original e tradução)
Finalidade	Estudo (análise terminológica)

Para que a análise proposta fosse feita, primeiramente o *corpus* foi compilado por meio do *download* das legendas da série *Supernatural* pelo site *legendas.tv*.[2] Este site é alimentado por legendas de diversas séries e filmes postadas por tradutores não especializados. Para as legendas serem lidas pelo *WordSmith Tools* versão 5.0 (SCOTT, 2008) é preciso que estejam em um arquivo de texto simples, isto é, em *.txt*. Felizmente as legendas compiladas no formato *.srt* não precisam ser convertidas, uma vez que, abertas no bloco de notas, elas também são lidas pelo programa.

Segue um exemplo de arquivo em *.srt* no bloco de notas de um trecho das legendas da série:

[2] Disponíveis em: www.legendas.tv.


supernatural.s01e01.720p.bluray.x264-mac-HI.srt - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

```
00:24:42,522 --> 00:24:45,774
Does Jessica know the truth? Does she
know about things you've done?

312
00:24:45,942 --> 00:24:48,402
No. And she's not ever
going to know.

313
00:24:48,570 --> 00:24:51,947
Well, that's healthy.
You can pretend all you want, Sammy.

314
00:24:52,199 --> 00:24:54,742
But you're gonna have to
face up to who you really are.

315
00:24:55,243 --> 00:24:56,744
-Who's that?
-You're one of us.

316
00:24:56,912 --> 00:24:58,412
No. I'm not like you.

317
00:24:58,580 --> 00:25:01,248
-This is not going to be my life.
-You have a responsibility.

318
00:25:01,416 --> 00:25:04,084
To Dad? And his crusade?

319
00:25:04,711 --> 00:25:08,088
If it weren't for pictures, I wouldn't
even know what Mom looks like.

320
00:25:08,256 --> 00:25:12,384
And what difference would it make?
Even if we find what killed her...

321
00:25:12,552 --> 00:25:15,179
...Mom's gone,
and she isn't coming back.
```

FIGURA 2 - Trecho das legendas em *.srt* do primeiro episódio da primeira temporada de *Supernatural*.

Como se trata de um *corpus* bilíngue, temos um sub-*corpus* em português e um em inglês, os quais, separadamente, foram em seguida utilizados para a geração de uma lista de palavras no *WordSmith Tools*, chegando-se aos seguintes dados sobre a constituição do *corpus*:

QUADRO 3
Dimensões do *corpus*

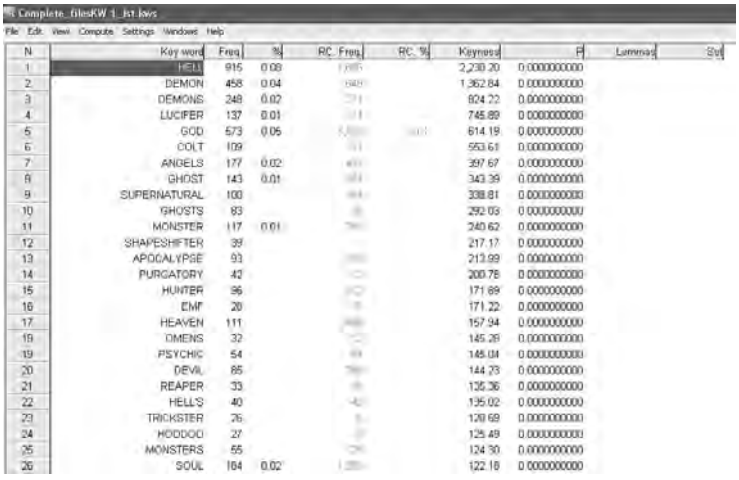
Sub-*corpus*	Itens (*tokens*)	Formas (*types*)
Inglês	1.122.505	15.030
Português	958.632	18.812
Total	2.081.137	—

O quadro separa itens e formas. Em cada sub-*corpus*, itens dão a totalidade de palavras que compõem as legendas em inglês e em português, enquanto formas indicam o número de vocábulos diferentes presentes nas legendas em cada língua. De acordo com Berber Sardinha (2004), nosso *corpus* de estudo seria classificado como médio-grande (entre 1 milhão a 10 milhões de palavras).

Percebe-se que, apesar de se tratar da tradução, o *corpus* em português tem 163.873 itens a menos que o *corpus* em inglês, ou seja, a partir de dados quantitativos já podemos inferir algumas considerações a respeito da tradução e equivalência linguística. É possível afirmar que um texto traduzido não terá a mesma quantidade de palavras que o seu original, visto que se trata de sistemas linguísticos diferentes. Essa é a primeira consideração a se fazer, porém há questões de tradução a serem consideradas, as quais, no momento, não serão discutidas.

Para que o levantamento dos candidatos a termos fosse feito, utilizou-se a ferramenta *KeyWords* do *WordSmith Tools*. Para que as listas de palavras-chave[3] sejam feitas, é necessário comparar uma lista de palavras do *corpus* de estudo e uma lista de palavras de um *corpus* de referência. Utilizou-se, então, como *corpus* de referência uma lista de palavras em inglês do *American National Corpus* (ANC) com 22.759.536 itens, visto que a série em análise apresenta a variação norte-americana da língua inglesa. Para o português, utilizou-se uma lista de palavras do *corpus* de referência Lácio-Ref com 9.602.849 itens. Em ambos os casos, os *corpora* de referência são bem maiores do que a proporção de cinco para um proposta por Berber Sardinha (2004) como o tamanho mínimo recomendado.

A seguir, apresentamos uma lista das palavras-chave em inglês e uma em português:



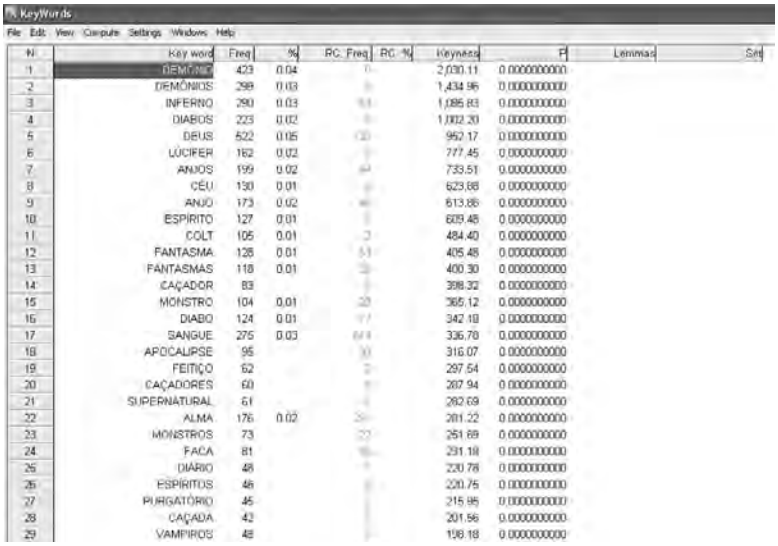
Complete_filesKW 1_list.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq	%	RC Freq	RC %	Keyness	P	Lemmas	Sort
1	HELL	915	0.09			2,230.20	0.000000000		
2	DEMON	458	0.04			1,362.84	0.000000000		
3	DEMONS	248	0.02			824.22	0.000000000		
4	LUCIFER	137	0.01			745.89	0.000000000		
5	GOD	573	0.05			614.19	0.000000000		
6	COLT	109				553.61	0.000000000		
7	ANGELS	177	0.02			397.67	0.000000000		
8	GHOST	143	0.01			343.39	0.000000000		
9	SUPERNATURAL	100				338.81	0.000000000		
10	GHOSTS	83				292.03	0.000000000		
11	MONSTER	117	0.01			240.62	0.000000000		
12	SHAPESHIFTER	39				217.17	0.000000000		
13	APOCALYPSE	93				213.99	0.000000000		
14	PURGATORY	42				200.78	0.000000000		
15	HUNTER	96				171.89	0.000000000		
16	EMF	28				171.22	0.000000000		
17	HEAVEN	111				157.94	0.000000000		
18	OMENS	32				145.28	0.000000000		
19	PSYCHIC	54				145.04	0.000000000		
20	DEVIL	85				144.73	0.000000000		
21	REAPER	33				135.36	0.000000000		
22	HELL'S	40				135.02	0.000000000		
23	TRICKSTER	26				128.69	0.000000000		
24	HOODOO	27				125.49	0.000000000		
25	MONSTERS	55				124.30	0.000000000		
26	SOUL	184	0.02			122.18	0.000000000		

FIGURA 3 - Palavras-chave em inglês

[3] Utilizamos também *Stoplists* em inglês e português para a exclusão de palavras gramaticais que não são de interesse para a nossa pesquisa.


KeyWords

File Edit View Compute Settings Windows Help

N	Key word	Freq	%	RC. Freq	RC. %	Keyness	P	Lemmas	Set
1	DEMÔNIO	423	0.04			2,030.11	0.000000000		
2	DEMÔNIOS	299	0.03			1,434.96	0.000000000		
3	INFERNO	290	0.03			1,086.83	0.000000000		
4	DIABOS	223	0.02			1,002.20	0.000000000		
5	DEUS	522	0.05			952.17	0.000000000		
6	LÚCIFER	162	0.02			777.45	0.000000000		
7	ANJOS	199	0.02			733.51	0.000000000		
8	CÉU	130	0.01			623.88	0.000000000		
9	ANJO	173	0.02			613.86	0.000000000		
10	ESPÍRITO	127	0.01			609.48	0.000000000		
11	COLT	105	0.01			484.40	0.000000000		
12	FANTASMA	126	0.01			405.48	0.000000000		
13	FANTASMAS	118	0.01			400.30	0.000000000		
14	CAÇADOR	83				398.32	0.000000000		
15	MONSTRO	104	0.01			365.12	0.000000000		
16	DIABO	124	0.01			342.18	0.000000000		
17	SANGUE	275	0.03			336.78	0.000000000		
18	APOCALIPSE	95				316.07	0.000000000		
19	FEITIÇO	62				297.54	0.000000000		
20	CAÇADORES	60				287.94	0.000000000		
21	SUPERNATURAL	61				282.69	0.000000000		
22	ALMA	176	0.02			281.22	0.000000000		
23	MONSTROS	73				261.69	0.000000000		
24	FACA	81				231.18	0.000000000		
25	DIÁRIO	48				230.78	0.000000000		
26	ESPÍRITOS	46				220.75	0.000000000		
27	PURGATÓRIO	45				215.95	0.000000000		
28	CAÇADA	42				201.56	0.000000000		
29	VAMPIROS	48				198.18	0.000000000		

FIGURA 4 - Palavras-chave em português

É por meio das palavras-chave que será possível identificar os candidatos a termos. A partir dessa identificação feita nas duas línguas, é necessário verificar quais termos estão presentes nas duas listas, a fim de se estabelecer quais deles são equivalentes. A seguir, são mostradas as palavras-chave a partir de uma planilha feita no Excel, seguindo a ordem de chavicidade[4] dos termos:

[4] O quanto a palavra é chave dentro do *corpus*.

PLANILHA 1
Relação de alguns candidatos a termos equivalentes em inglês e português

Ordem	Inglês	Ordem	Português
1	HELL	3	INFERNO
2	DEMON	1	DEMÔNIO
4	LUCIFER	6	LÚCIFER
5	GOD	5	DEUS
6	COLT	11	COLT
8	GHOST	12	FANTASMA
11	MONSTER	15	MONSTRO
12	SHAPESHIFTER	49	METAMORFO
13	APOCALYPSE	18	APOCALIPSE
14	PURGATORY	27	PURGATÓRIO
15	HUNTER	14	CAÇADOR
16	EMF	153	EMF
17	HEAVEN	8	CÉU
20	DEVIL	16	DIABO
21	REAPER	250	CEIFADOR
23	TRICKSTER	287	TRICKSTER
24	HOODOO	913	HOODOO
26	SOUL	22	ALMA

Tomando as palavras-chave como palavra de busca, utilizamos a ferramenta *Concord* para visualizarmos as linhas de concordâncias. Estas mostram o termo em destaque em seu contexto de uso, e é a partir desse contexto que as definições para a elaboração do glossário serão criadas. É importante lembrar que talvez nem todos os termos sejam adicionados ao banco de dados pela possível falta de contexto que possibilite a construção de uma definição clara e objetiva.

Em seguida, apresentamos um exemplo das linhas de concordância do termo *inferno*:

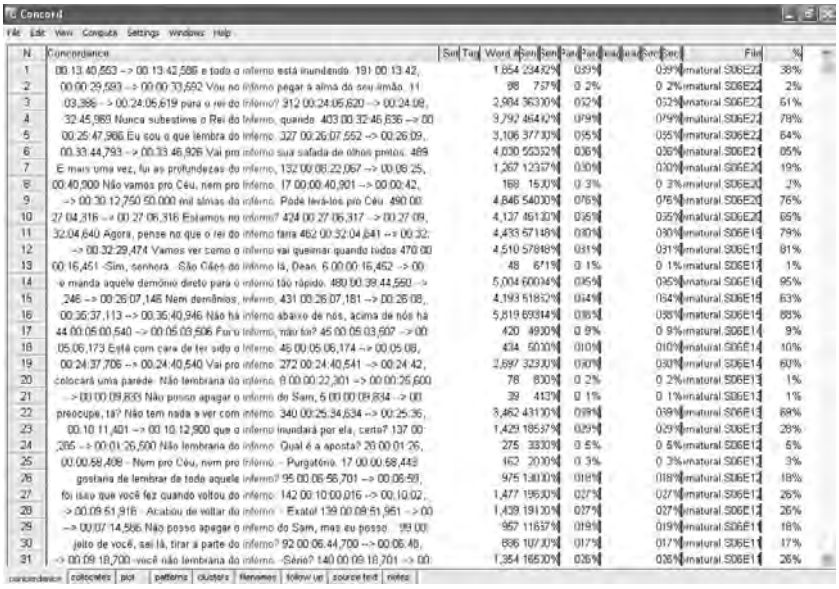


FIGURA 5 - Linhas de concordância do termo *inferno*

Visto que um dos nossos objetivos é construir um banco de dados, para que um glossário seja gerado, as concordâncias fornecerão os contextos necessários para que as definições sejam criadas.

Os contextos podem ser de três tipos.

> O *contexto associativo* apresenta o termo como pertinente ao tema objeto da pesquisa, mas não indica os traços conceptuais específicos destes termos, [...] Já os *contextos explicativos* apresentam alguns traços conceptuais pertinentes específicos do termo sob observação, freqüentemente relativos à materialidade, finalidade, funcionamento e similares. [...] Talvez mais desejáveis, mas certamente menos encontradiços, os *contextos definitórios* proporcionam um conjunto completo dos traços conceptuais distintivos do termo. Tal distintividade, no entanto, representa freqüentemente um certo nível de abstração, sem indícios claros da gama efetiva de usos em situação do termo (AUBERT, 1996 apud FROMM, 2007).

A partir de algumas observações já feitas, nota-se que uma característica do *corpus* em estudo é a grande quantidade de contextos definitórios. Muitos dos termos ao serem utilizados são definidos

explicitamente pelos personagens da série, visto que, ao lidarem com criaturas desconhecidas, eles buscam pelas características das mesmas, fornecendo vários traços distintivos importantes para a definição do termo.

Para a inserção das informações pertinentes aos termos, é necessário preencher uma ficha terminológica. A seguir, apresentamos parte dessa ficha:

FIGURA 6 - Trecho da ficha terminológica do VoTec (FROMM, 2007)

Os campos, a serem preenchidos com as informações de cada termo proveniente do *corpus*, podem ser detalhados da seguinte forma:

- Dados: referem-se à categoria gramatical, número, gênero, siglas, acrônimos, variações morfossintáticas, posição na ordem de freqüência do *corpus*, número de ocorrências do termo;
- Traços Distintivos: campo utilizado para a discriminação das características principais extraídas dos contextos que ajudarão na criação da definição final do termo;
- Semântica: é o campo onde se explicitam as relações de hiperonímia, hiponímia, co-hiponímia, antonímia e sinonímia, bem como informações de se o termo já foi dicionarizado, em qual dicionário e qual foi a definição apresentada;

- Termo equivalente: aqui são reconhecidos os termos equivalentes, o que faz do vocabulário um vocabulário bilíngue;
- Termos remissivos: nesse campo inserem-se os termos que de alguma forma estabelecem uma relação semântica com outros termos;
- Informações enciclopédicas: campo utilizado para a inserção de informações provenientes de outras fontes.

Para a construção do banco de dados, utilizaremos a base informatizada VoTec desenvolvida por Fromm (2007). A seguir, apresenta-se a tela da área de consulta de termos disponível gratuitamente na Internet pelo seguinte endereço: <www.ic.voteconline.com.br>.

FIGURA 7 - Página de consulta do VoTec (FROMM, 2007)

Percebe-se que antes da consulta é preciso selecionar as áreas nas quais se quer buscar um termo como as exemplificadas acima: Terminologia em Ficção – Sobrenatural. Observa-se também que a página disponibiliza o termo em português, bem como a sua definição em português e seu equivalente em inglês definido em inglês. Isso demonstra que não se trata apenas de um vocabulário bilíngue, com um termo e sua tradução. Ele vai além, possibilitando a elaboração de definições nas duas línguas e comparando-as.

6 Próximos passos

Além dos procedimentos exemplificados acima, pretende-se realizar uma análise contrastiva levando em conta a teoria das modalidades de tradução proposta por Aubert (1998), na qual as traduções das legendas serão comparadas com as originais em inglês para se verificar a tradução dos termos e prováveis estratégias utilizadas pelos tradutores, bem como finalizar a inserção dos termos no banco de dados.

Portanto, a partir do exposto, apresentamos uma visão geral dos aspectos teóricos e metodológicos utilizados na realização desta pesquisa, bem como a caracterização do que chamamos de vocabulário do horror. Verifica-se, então, a sua aplicabilidade tanto para os estudos terminológicos e terminográficos, quanto para os estudos em tradução e para a análise de *corpora* eletrônicos especializados.

Referências

ALMEIDA, G. M. B. Fazer Terminologia é fazer Linguística. In: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. (Org.). *Linguagens especializadas em corpora*: modos de dizer e interfaces de pesquisa. Porto Alegre: EDIPUCRS, 2010.

AUBERT, F. H. Modalidades de tradução: teoria e resultados. *TradTerm*. São Paulo, v. 1, n. 5, p. 99-128, 1998.

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

BERBER SARDINHA, T. *Pesquisa em Linguística de Corpus com WordSmith Tools*. Campinas: Mercado de Letras, 2009.

BOWKER, L.; PEARSON, J. *Working with specialized language*: a practical guide to using corpora. London/New York: Routledge, 2002.

FINATTO, M. J. B.; KRIEGER, M. da G. *Introdução à terminologia*: teoria e prática. São Paulo: Contexto, 2004.

FROMM, G. *VoTec*: a construção de vocabulários técnicos eletrônicos para aprendizes de tradução. 2007. 215f. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês) - Departamento de Letras Modernas, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

NOLET, D.; PAVEL, S. *Manual de Terminologia*. Trad. Enilde Faulstich. Canadá: Departamento de Tradução, 2002.

OLOHAN, M. *Introducing corpora in translation studies*. London/New York: Routledge, 2004.

SCOTT, M. *WordSmith Tools*. Versão 5. Oxford: Oxford University Press, 2008.

TODOROV, T. *Introdução à literatura fantástica*. Trad. Maria Clara Correa Castello. São Paulo: Perspectiva, 2008.

WELKER, H. A. *Dicionários* - uma pequena introdução à lexicografia. 2. ed. Brasília: Thesaurus, 2004.

Um estudo de corpus das metáforas do conceito *sociedade* em alemão e em português

Emanuela G. Costa[1]

RESUMO: O objetivo deste trabalho foi analisar os padrões linguísticos que representam os mapamentos conceituais referentes ao conceito *sociedade*. Tendo em vista, que a Linguistcia de Corpus pode oferecer poderosas ferramentas para a análise de dados linguísticos, foram criados dois corpora, um em alemão e outro em português para se verificar como as metáforas linguísticas nos permitem acessar o sistema conceitual.

PALAVRAS-CHAVE: Metáfora, sociedade, Linguística de Corpus.

ABSTRACT: The goal of this project was to analyze the linguistic patterns that represent the conceptual mapping for the concept of *society*. Given that Corpus Linguistics can provide powerful tools for analyzing linguistics data, two corpora were built, one in German and another in Portuguese to verify how linguistic metaphors can allows us to access the conceptual system.

KEYWORDS: Metaphor, society, Corpus Linguistics.

1 Introdução

Desde o lançamento do livro *Metaphor we live by* (1980), George Lakoff e Mark Jonhson inauguram a linha de pesquisa da teoria conceitual da metáfora, na qual a metáfora deixa de ser vista apenas como uma figura da retórica, mas assume o papel de ser parte do nosso sistema conceitual, ou seja, o sistema orientador do nosso pensamento.

[1] COSTA, Emanuela G., mestranda no Programa de Pós- Graduação de Estudos Linguísticos da FALE/UFMG emanuela.costa@gmail.com.

No livro, os autores demonstram que o uso das metáforas está relacionado ao modo como entendemos vários conceitos existentes na língua, pois a operação cognitiva que se processa é o emprego do domínio-fonte, mais experiencial, com o domínio-alvo, mais abstrato, assim entendemos o conceito mais abstrato, por meio de um mais concreto (*i.e.*, B por meio de A), o que demanda o uso de uma metáfora (LIMA, 2001). Desse modo os autores apresentam uma distinção entre *metáfora conceitual* e as *metáforas lingüísticas*. A metáfora conceitual é o resultado entre mapeamentos de domínios, enquanto que as metáforas linguísticas são itens linguísticos individuais, resultantes do mapeamento conceitual. Como podemos demonstrar no exemplo (1):

(1) RAIVA É FOGO

a) Aqueles eram comentários *inflamados*.

b) Ela estava *cuspindo fogo*.

c) Ele foi *consumido* pela *raiva*.

(STEFANOWITSCH & GRIES, 2006:65)

Desse modo, percebemos que estudos a partir da abordagem conceitual da metáfora, aspiram, sobretudo, descrever os mapeamentos conceituais, e para chegar às conceitualizações é preciso utilizar exemplo da língua. Desse modo, a maior parte dos exemplos era captada introspectivamente, ou coletados de apenas um a autor.

Com o advento da Linguística de Corpus, que nos últimos 15 anos tem se tornado uma das mais fortes metodologias de análises empíricas da Linguística, por oferecer uma base de dados autênticos da língua em uso, um estudo da metáfora baseada em corpus pode trazer novos ganhos, já que implica em demonstrar empiricamente quais padrões linguísticos representam algumas metáforas conceituais.

Tendo em vista o estudo desenvolvido em Schröder 2009, sobre a construção metafórica do termo *sociedade*, português e alemão, concluiu-se que havia diferença na conceptualização desse conceito nas duas línguas.

O estudo revelou uma tendência do corpus em alemão em metaforizar *sociedade* por meio de esquemas imagéticos misturados e dinamizados, além disso, as metáforas conceituais encerraram em muitos casos, como domínio fonte os conceitos de negócios, prédios,

jogos e observação. O corpus em português, por outro lado, tendia a ser motivado pelos domínios-fonte da família, da guerra, flora e estágio, além de ter sido mais personificado. Como nos exemplo (2).

(2) SOCIEDADE É UM CORPO

a) "saindo da *barriga* da família passando para a *barriga* da sociedade, né"

b) "Com o road map nas *mãos*, o caminho está dado." (Schröder 2009:121)

Em alemão temos o exemplo (3), representa outra metáfora conceitual.

(3) A SOCIEDADE É UMA EMPRESA

a) "Werte*haushalt* der deutschen Gesellschaft"
(*orçamento* de valores da sociedade alemã)

b) "Generationen-*Buchhaltung*"
(geração de *contabilidade*) (Schröder 2009: 124)

Logo, para se falar de um termo puramente abstrato com *sociedade*, a mente faz elaborações complexas; assim, propõem-se neste estudo padrões linguísticos que envolvem o conceito *sociedade*.

2 Corpora e metodologia

Este estudo tem como objetivo descrever um conjunto de dados linguísticos, sob uma perspectiva comparativa em duas línguas; português e alemão. Ambos os corpora são formados por dados digitais. Os dados pertencem ao gênero mídia online, formados por 145 textos em português, retirados da revista *Carta Capital online* e por 117 textos em alemão retirados da revista *Der Spiegel online.*

Logo, podemos defini-los como:

Tipologia: escrito

Tempo: Sincrônico e contemporâneo. Recolhidos entre os dias 30/05/11 até o dia 30/06/11

Seleção: Amostragem

Conteúdo: Especializados. Os dados foram recolhidos apenas da coluna *política* de ambas as revistas.

Gênero: jornalístico, online.

Autores: Os textos foram produzidos por falantes nativos.

Finalidade: Esse corpus tem como objetivo retratar a representação do lexema *sociedade* metaforicamente.

Uma das maiores dificuldades na produção de corpora comparativos é sempre em relação ao seu tamanho, já que a qualidade das informações está diretamente ligada a esse fato. Nesta pesquisa foi delineado um corpus em português formado por 145 textos, gerando o total de 68.828 token[2] e 10.696 types[3] e um corpus em alemão formado por 117 textos, gerando o total de 75.547 token e 12.967 types. Percebemos que apesar do corpus em alemão conter um numero menor de textos, ele ficou maior do que aquele feito em português. Isso ocorreu porque a coluna política da revista *Der Spiegel* era formada por texto que continham em média 6kB[4] de informação, enquanto os textos em português da revista *Carta Capital* contavam com uma média de 3kB de informação. Como a intenção era analisar as ocorrência em ambientes muito próximos, tivemos que diminuir o numero de textos em alemão para conseguir contrastar com os textos em português.

Segundo Sardinha (2004) ambos os corpora estão delimitados para serem considerados como um pequeno, já que contam com um número inferior a 80 mil palavras. No entanto, como esses corpora visam servir de amostra para uma ocorrência específica, portanto, acredita-se que eles cumprem esta função inicial.

Os dados foram recolhidos dos sites das revistas e transferidos para o programa *bloco de notas* do Windows, em seguida foram analisados por meio de um programa chamado *Antcon*, que pode nos dar o numero total de types, token, e as linhas de concordância que integravam o lexema *sociedade/ Gesellschaft*.O lexema *sociedade* ocorreu na posição 213, com 35 aparições ao longo do corpus.

[2] Token: número total de lexemas.

[3] Types: número total da variação de lexemas.

[4] kB: kilobyte.

Exemplo (4):
a) os que criar uma sociedade da igualdade substancial. A produ
b) os e riquezas da sociedade. Para mudar isso, será preciso u
c) ai ser a base da sociedade do futuro. E não podemos ter certeza d
d) s em detrimento dos reais problemas que assolam a sociedade americana,
e) o essa que nossa sociedade se acomodou para afrontar seus probl
f) Nossa sociedade deverá se confrontar

Já o lexema *Gesellschaft* ocorreu na posição 496, com 19 aparições ao longo do corpus.

Exemplo (5):
a) n Platz in der Mitte der Gesellschaft zu finden. Dennoch warf
b) werdenden Gesellschaft gar nicht anders sein
d) welcher Grenze kann eine Gesellschaft um des Wohlstands wi
e) dliche Nebenprodukt einer Gesellschaft, in der ü

Destarte, percebemos que apesar de o corpus em alemão ser maior do que aquele em português, a frequência do lexema de análise em português foi maior.

3 Resultados

3.1 Português

Após abrirmos as linhas de concordâncias percebemos que das 35 ocorrências, a maioria ocorre com a metáfora conceitual:

(6) SOCIEDADE É UM CORPO.
a) "...Nossa sociedade deverá se confronta..."
b) "...de porque a sociedade clama..."
c) "...desta casa legislativa perante a sociedade..."
d) "A sociedade fica mais tolerante e passa a reconhecer..."
e) "...tanta confiança da sociedade, merecem, igualmente,..."

A primeira observação dos dados permite que percebamos a metaforizarão do lexema *sociedade* como CORPO, devido à existência

de verbos de ação, tais como: se confrontar, clamar, ficar mais tolerante, ter confiança; que humanizam o conceito sociedade. Além disso, podemos observar os elementos que ocorrem com o termo analisado. Promovendo a seguinte regra:

DETERMINANTE+ SOCIEDADE+ VERBO

Em seguida observamos os dados da metáfora conceitual,

(7) A SOCIEDADE É UM CAMINHO
a) ser a base da sociedade do futuro."
b) passos em direção a uma sociedade mais justa,.."
c) despertar na sociedade, principalmente LGBT..."

Gerando a seguinte regra:

PREPOSIÇÃO EM+ (DET.)SOCIEDADE+

Finalmente a ultima metáfora conceitual que apareceu nos dados com apenas três ocorrências

(8) SOCIEDADE É UMA FONTE.
a) "...e riquezas da sociedade."
b) "...os problemas da nossa sociedade.."
c) "...para a sociedade civil numa fonte..."

3.2 Alemão

O lexema *sociedade/ Gesellschaft* ocorreu 19 vezes no corpus. Vejamos alguns dados analisados. Em primeiro lugar, houve duas ocorrências como nomes próprios, isto é, sem nenhuma relação metafórica.

Exemplo (9):
"...die Max-Planck-Gesellschaft, die Leibniz-Gemeinschaf..."
"...und die Fraunhofer-Gesellschaft gehören, stellte sich…"

Com o esquema imagético

(10) CENTRO-PERIFERIA.

a) in der *Mitte der* Gesellschaft zu finden/ para encontrar *no meio* da sociedade

(11) A SOCIEDADE É UM CORPO

a) "...*älter werdenden* Gesellschaft gar nicht anders sei/ a envelhecida sociedade não pode nada mais

Regra: ADJETIVO+SOCIEDADE+…

A SOCIEDADE É UMA BOMBA

a) "Ich denke, die Gesellschaft wird *explodieren*/ Eu acho, que a sociedade *explodirá*

Regra: DETERMINANTE+ SOCIEDADE+ VERBO

4 Conclusão

Após explicar os resultados encontrados nos corpora selecionados, percebemos mais uma vez, que para se falar de conceitos puramente abstratos, fazemos uso de estruturas pré-existentes em nossas línguas. Além disso, percebemos como essas estruturas aparecem frequentemente na nossa linguagem diária.

Contudo, ficou a impressão de que outras conclusões poderiam ser encontradas a partir dos mesmos corpora, uma vez que essa pesquisa se baseou em apenas um determinado lexema, *sociedade*, e por isso os resultados encontrados foram tímidos em relação ao potencial de análise.

Portanto, acredita-se que essa mesma pesquisa, a partir do mesmo corpus, poderia ser expandida, com o uso de alguns sinônimos do lexema estudado, tais como *nação* ou mesmo os *nomes dos países*, nos quais os textos se inserem, nesse caso: Brasil e Alemanha. Consequentemente, acredita-se que a pesquisa conseguiu, ainda que com resultado inicial, demonstrar um pouco do potencial do uso da Linguística de Corpus, como metodologia de análise na Teoria Cognitiva da Metáfora.

Referências

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

DEIGNAN, Alice.; POTTER, Liz. *A corpus study of metaphors and metonyms in English and Italian*. Birmingham, Uk: Elsevier, 2000.

LAKOFF, G.; JOHNSON, M. *Metaphors we live by*. Chicago: The University of Chicago Press, 1980.

LIMA, Paula Lenz. Emergência e natureza da metáfora primária: desejar é ter fome. *Caderno de Estudos Linguísticos*, Campinas, jun. 2001. Disponível em: <http://www.lafape.iel.unicamp.br/Publicacoes/paula_gibbs_edson.pdf>. Acesso em: 20 ago. 2008.

MARTELOTTA, Mario E. *Manual de Linguística*. São Paulo: Editora Moderna, 2002.

SARDINHA, Tony B. *Linguística de Corpus*. Barueri: Editora Manole, 2004.

SARDINHA, Tony B. *Metáfora*. São Paulo: Parábola Editorial, 2008.

SCHRÖDER, Ulrike. A construção metafórica do conceito 'sociedade' em perspectiva comparativa. *Revista Pandaemonium Germanicum*, São Paulo, v. 2009.2, p. 105-141. Disponível em: <http://www.fflch.usp.br/dlm/alemao/pandaemoniumgermanicum/site/images/pdf/ed2009.2/06_Schroeder_Ulrike.pdf>. Acesso em: 30 jun. 2011.

STEFANOWITSCH, Anatol; GRIES, Stephan. *Corpus-based approaches to metaphor and metonymy*. [Trends in Linguistics Studies and Monographs 171.] Berlin & Nova York: Mouton de Gruyter, 2006.

Significados existenciais no português brasileiro: um estudo contrastivo em textos traduzidos e não traduzidos

Kícila Ferreguetti[1]
Adriana Pagano[2]
Giacomo Figueredo[3]

RESUMO: O presente trabalho apresenta resultados preliminares da pesquisa sobre orações existenciais realizada pelo grupo de pesquisa *Modelagem sistêmico-funcional da tradução e da produção textual multilíngue*, desenvolvido pelo LETRA/FALE/UFMG. A pesquisa tem como objetivo identificar e analisar a ocorrências de orações existenciais no corpus monolíngue CALIBRA (Catálogo da Língua Brasileira), bem como em um corpus paralelo bilíngue (composto dos originais e das traduções para o italiano de duas obras da autora Clarice Lispector), visando contribuir para a descrição sistêmico-funcional da realização dos significados existenciais no português brasileiro e para as pesquisas no âmbito dos estudos da tradução. As orações existenciais são consideradas importantes para o discurso por realizarem gramaticalmente a existência e o acontecimento (HALLIDAY; MATTHIESSEN, 2004). A análise foi dividida em duas etapas, tendo em vista os dois corpus utilizados, e consistiu na busca e extração de linhas de concordâncias contendo os processos existenciais feita com o auxílio do software *WordSmith Tools* (Scott, 2007).

PALAVRAS-CHAVE: linguística sistêmico-funcional, orações existenciais, análise contrastiva, estudos da tradução.

[1] Mestranda em Estudos Linguísticos (Estudos da Tradução). Universidade Federal de Minas Gerais. kicilaferreguetti@yahoo.com.br

[2] Doutora em Letras. Professora Associada da Universidade Federal de Minas Gerais. apagano@ufmg.br.

[3] Doutor em Linguística Aplicada. Professor Adjunto de Linguística Aplicada. Universidade Federal de Ouro Preto. giacomojakob@yahoo.ca.

ABSTRACT: This paper presents the preliminary results of a study on existential clauses carried out by the research group *Modelagem sistêmico-funcional da tradução e da produção textual multilíngue* at LETRA/FALE/UFMG. This research aims at identifying and analyzing the occurrences of existential clauses in a monolingual corpus (CALIBRA) as well as in a bilingual parallel corpus (a compilation of two books of Brazilian author Clarisse Lispector and their translations to Italian). The findings are expected to contribute to a systemic-functinal description of Brazilian Portuguese and to research within the discipline of Translation Studies researches. The analysis comprised two stages which consisted of querying the corpora and extracting concordance lines with existential clauses using the software *WordSmith Tools* (Scott, 2007).

KEYWORDS: systemic functional linguistics, existential clauses, contrastive analysis, translation studies.

1 Introdução

O presente trabalho apresenta os resultados preliminares da pesquisa sobre orações existenciais realizada pelo grupo de pesquisa *Modelagem sistêmico-funcional da tradução e da produção textual multilíngue*, desenvolvido no Laboratório Experimental de Tradução (LETRA) da Faculdade de Letras/UFMG.

A pesquisa visa contribuir para a descrição sistêmico-funcional do português brasileiro iniciada em Figueredo (2011) e, também, para as pesquisas no campo disciplinar dos estudos da tradução. Sendo assim, possui um duplo objetivo: identificar e analisar a ocorrências de orações existenciais no corpus monolíngue CALIBRA (Catálogo da Língua Brasileira), bem como em um corpus paralelo bilíngue (composto dos originais e das traduções para o italiano de duas obras da autora Clarice Lispector).

Segundo Halliday e Matthiessen (2004), os seres humanos utilizam a linguagem para construir ou representar as suas experiências no mundo. A metafunção responsável por realizar essa função é a experiencial, através do sistema de Transitividade. Essas experiências, por sua vez, correspondem aos eventos que estão acontecendo no mundo e que o sistema de Transitividade organiza e agrupa na oração na forma de uma figura composta por um processo, participantes e circunstâncias.

Os processos são considerados os principais componentes do sistema de Transitividade porque são eles os responsáveis por construir a figura que irá representar a experiência na oração. Possuem sempre pelo menos um participante ligado a eles e podem ser acompanhados ou não de uma circunstância (de tempo, modo, lugar, entre outras). Dividem-se em: processos materiais, mentais, relacionais, verbais, comportamentais e existenciais (HALLIDAY; MATTHIESSEN, 2004).

FIGURA 1
A gramática da experiência: tipos de processos na língua inglesa
Fonte: Traduzida e adaptada de HALLIDAY; MATTHIESSEN (2004, p. 172) para a descrição da língua portuguesa.

Como ilustrado pela Figura 1 acima, os processos materiais representam aquilo que está sendo feito ou acontecendo no mundo, isto é, as mudanças que ocorrem no mundo físico. Os mentais, por outro lado, constroem aquilo que está acontecendo no âmbito da consciência das pessoas, ou seja, aquilo que se passa no interior de

suas mentes, suas emoções e suas percepções do mundo. Já os processos relacionais constroem relações, que podem ser de identidade ou de atribuição. Os verbais, por sua vez, são os processos responsáveis por realizar ou relatar aquilo que é ou foi dito. Finalmente, os existenciais, foco deste trabalho, como o próprio nome indica, constroem a existência (HALLIDAY; MATTHIESSEN, 2004).

Ainda com relação à Figura 1, é possível observar que os processos existenciais estão localizados entre os processos relacionais e os materiais, mais especificamente entre os relacionais de atribuição e os materiais de acontecimento. Em outras palavras, os processos existenciais estão localizados no limite do mundo das relações abstratas e do mundo físico.

As orações existenciais, por sua vez, são caracterizadas pela presença de um processo existencial e de um participante (Existente), que pode ser desde uma pessoa até um evento. Além disso, as orações existenciais também podem possuir uma circunstância de tempo e lugar (HALLIDAY; MATTHIESSEN, 2004). Os exemplos 1 e 2 abaixo ilustram este tipo de oração:

Exemplo 1:

Havia	**lugares pobres e ricos que precisavam dela.**
Processo existencial	Existente

Exemplo 2:

Havia	**no chão**	**caroços secos**
Processo existencial	Circunstância de lugar	Existente

Ainda segundo os autores, as orações existenciais, ainda que não possuam uma frequência de ocorrência alta no discurso (se comparadas às demais orações), são relevantes na construção dos diferentes tipos de texto, uma vez que realizam gramaticalmente a existência e o acontecimento. No entanto, os processos e orações existenciais não têm sido objeto destacado nas pesquisas realizadas no âmbito dos estudos sistêmicos funcionais.

2 Metodologia

Os dados apresentados e analisados neste trabalho foram extraídos de dois corpus: um monolíngue, o CALIBRA (Catálogo da Língua Brasileira) e outro paralelo bilíngue (português/italiano). O CALIBRA, desenvolvido em parceria pelo Laboratório Experimental de Tradução (LETRA), da Faculdade de Letras da UFMG, e pelo Instituto de Ciências Humanas e Sociais (ICHS), da UFOP, é um corpus de um milhão de palavras, subdividido em oito subcorpora (capacitar, compartilhar, explicar, explorar, fazer, recomendar, recriar e relatar). Cada subcorpus representa um tipo de processo sócio-semiótico e é composto por monólogos e diálogos tanto escritos como falados. O corpus paralelo bilíngue, por sua vez, é composto pelos originais em português e suas traduções para italiano das obras *A hora da estrela* e *Laços de Família* da autora Clarice Lispector.

Sendo assim, a pesquisa foi dividida em duas etapas. A primeira consistiu na identificação e análise das ocorrências de orações existenciais nos textos pertencentes a dois subcorpora do CALIBRA: o Recomendar e o Relatar. Já segunda etapa consistiu na identificação e analise das ocorrências das orações existenciais no corpus paralelo bilíngue.

A análise, em ambas as etapas, foi realizada com o auxílio do software *WordSmith Tools* (Scott, 2007) e de suas ferramentas: *Wordlist*, *Concord* e *Aligner*. Primeiramente, através da ferramenta *Wordlist*, foi gerada uma lista de palavras com todas as palavras que ocorrem nos dois corpus, visando a identificação dos verbos lexicais que prototipicamente realizam a existência. O Quadro 1 abaixo, apresenta as ocorrências encontradas em ambos os *corpora*.

QUADRO1
Lemas de verbos lexicais passíveis de realizar processos existenciais

Acabar	Dar	Ficar	Permanecer	Sair
Acontecer	Decorrer	Haver	Persistir	Seguir
Aparecer	Desaparecer	Iniciar	Pintar	Ser
Apontar	Durar	Ir	Prevalecer	Sobrar
Brotar	Encontrar	Jazer	Resistir	Sumir
Chegar	Existir	Manifestar	Restar	Surgir
Começar	Extinguir	Morrer	Rolar	Ter
Comparecer	Faltar	Nascer	Romper	Vir

Em seguida, utilizando a ferramenta *Concord*, foi realizada uma busca pelos radicais dos verbos lexicais listados no Quadro 1 acima, visando incluir as ocorrências com flexões de tempo e número. Apenas as linhas de concordância com as ocorrências desses verbos lexicais realizando processos existenciais foram selecionadas. A anotação dos dados, no entanto, foi feita manualmente.

No que diz respeito ao corpus paralelo bilíngue, é importante ressaltar que, para este trabalho, foram buscados apenas os verbos lexicais que realizam processos existenciais com maior frequência, conforme dados obtidos no corpus monolíngue, isto é, "existir", "haver" e "ter", juntamente com os seus equivalentes em italiano "*esistere*" e "*esserci*". Além disso, os textos (originais e suas respectivas traduções) foram alinhados com o auxílio da ferramenta *Aligner*, para que as ocorrências identificadas pudessem sem comparadas. O objetivo era verificar não só como as orações existenciais presentes nos originais em português foram traduzidas para o italiano, como também se ocorreram mudanças.

3 Análise dos dados

3.1 As orações existenciais no corpus monolíngue CALIBRA

A análise das ocorrências das orações existenciais no subcorpus Recomentar revelou a ocorrência de 22 orações existenciais, sendo realizadas por 4 tipos de verbos lexicais diferentes, como pode ser observado a partir da Figura 2, a seguir, que apresenta também alguns exemplos extraídos do corpus:

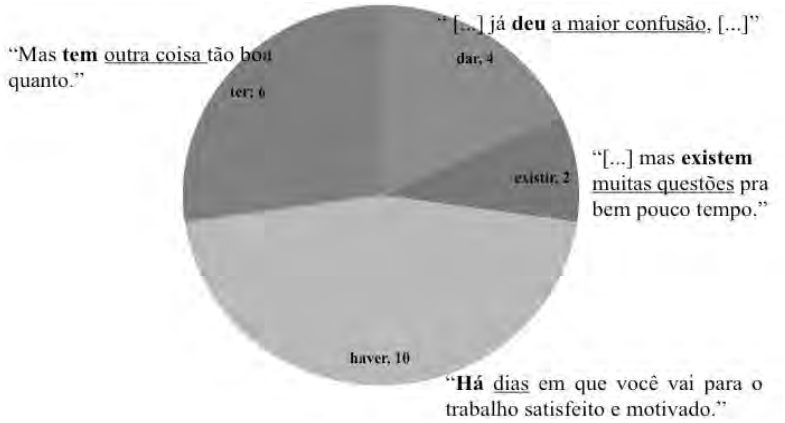


FIGURA 2
As orações existenciais no processo Recomendar
Legenda: Os processos existenciais aparecem em **negrito**; os participantes (existentes), sublinhados

A partir da análise da Figura 2 é possível observar que os verbos lexicais "haver" e "ter" são os que realizam processos existenciais com maior frequência no subcorpus relatar: 10 e 6 ocorrências respectivamente. É importante ressaltar, também, que, além dos verbos lexicais que comumente realizam a existência em português (haver, ter e existir), foram identificadas no subcorpus 4 ocorrências do verbo lexical "dar" realizando essa função, sendo que estas foram até mais frequentes do que as do verbo "existir".

Já no subcorpus Relatar foram verificadas a ocorrência de um número consideravelmente maior de orações existenciais: 368 no total, sendo realizadas por 18 tipos de verbos lexicais diferentes, como ilustrado pela Figura 3, a seguir, que também oferece exemplos das ocorrências mais frequentes no corpus:

FIGURA 3
As orações existenciais no processo Relatar
Legenda: Os processos existenciais aparecem em **negrito**; os participantes (existentes), sublinhados, e as circunstâncias, em *itálico.*

A análise dos dados apresentados na Figura 3 acima, revela que "haver", "ter" e existir são os verbos lexicais que realizam processos existenciais com maior frequência no subcorpus Relatar: 156, 102 e 56 ocorrências respectivamente. Além disso, foram encontrados outros 15 tipos diferentes de verbos lexicais realizando significados existenciais. Dentre eles, os mais frequentes foram: acontecer (11), aparecer (11), começar (7) e morrer (6).

Comparando os resultados obtidos na análise dos dois subcorpora é possível estabelecer que os verbos lexicais realizando processos existenciais mais frequentes são "haver", "ter", "existir", "acontecer" e "aparecer", sendo que "haver", realizando a existência, é o processo mais frequente em ambos os subcorpora, seguido de "ter".

Além disso, é possível observar uma ocorrência significativamente maior de orações existenciais no subcorpus Relatar do que no subcorpus Recomendar, uma vez que foram verificadas 368 ocorrências no primeiro contra apenas 22 ocorrências no segundo.

O subcorpus Relatar possui ainda uma maior variedade de verbos lexicais realizando significados existenciais do que o Recomendar. Foram identificadas 18 ocorrências de verbos lexicais realizando processos existenciais no subcorpus Relatar, enquanto o Recomendar apresentou apenas 4.

3.2 As orações existenciais no corpus paralelo bilíngue

No que diz respeito ao corpus paralelo bilíngue, como mencionado anteriormente, a análise das ocorrências de orações existenciais nos textos originais (as obras *A hora da estrela* e *Laços de Família* da autora Clarice Lispector) levou em consideração apenas os três verbos lexicais realizadores processos existenciais mais frequentes: "haver", "ter" e "existir". Os dados revelaram que os três ocorreram 138 vezes no corpus, distribuídos da seguinte forma:

TABELA1
Distribuição das orações existenciais em *Laços de Família* e *A hora da estrela*

Verbos realizadores de processos existenciais	**Número de ocorrências em *Laços de Família***	**Número de ocorrências em *A hora da Estrela***	**Número de ocorrências no Corpus**
Existir	9	16	25
Haver	**45**	**60**	105
Ter	0	8	8
Total	54	84	**138**

A partir da análise da Tabela 1, acima, é possível observar que a obra *A hora da Estrela* apresenta mais orações existenciais do que a obra *Laços de Família*. São 84 ocorrências de orações existenciais na primeira contra 54 ocorrências na segunda. Os dados revelam ainda que, quando se compara o número de ocorrências dos três verbos lexicais buscados, "haver" é o mais frequente em ambos os livros.

Já no que diz respeito às traduções, a busca por orações existenciais também enfocou os verbos lexicais realizadores de processos existenciais mais frequentes, no caso, o verbo lexical "*esistere*" (existir) e o verbo lexical "*esserci*" (o equivalente a "haver" e "ter" com o sentido de "existir" em italiano).

Sendo assim, no corpus como um todo, foram encontradas 33 ocorrências do verbo lexical "*esistere*" e 71 ocorrências do verbo lexical "*esserci*", totalizando 104 ocorrências, distribuídas da seguinte forma:

TABELA 2
Distribuição das orações existenciais em Legami familiari e L'ora della Stella

Verbos realizadores de processos existenciais	**Número de ocorrências em *Legami familiari***	**Número de ocorrências em *L'ora della Stella***	**Número de ocorrências no Corpus**
Esistere	10	23	33
Esserci	**17**	54	71
Total	**27**	77	**104**

A análise dos dados presentes na Tabela 2 revela que as traduções possuem menos ocorrências de orações existenciais (104) do que os originais (138), sendo que a maior diferença é observada na tradução de *Laços de Família*, que possui apenas a metade das orações existenciais observadas no original. Em *Legami familiari*, no total, foram identificadas 27 ocorrências, enquanto o original possui 54.

Por outro lado, é possível observar que ambas as traduções possuem mais ocorrências do verbo lexical "*esistere*" (existir) do que os originais, tanto individualmente quanto no corpus como um todo. São no total 33 ocorrências no corpus traduzido contra 25 ocorrências no corpus original.

No entanto, o que mais chama a atenção é a disparidade identificada entre as ocorrências do verbo lexical *esserci* (haver e ter com sentido de existir em italiano), quando comparadas as ocorrências de "haver" e "ter" nos originais. Nas traduções, no total, foram encontradas 71 ocorrências, enquanto os originais apresentam 113.

Essa disparidade se torna ainda mais relevante quando se considera os livros isoladamente. Novamente, a tradução de *Laços de Família* possui menos significados existenciais sendo realizados pelo verbo lexical *esserci*, equivalente do verbo lexical "haver", que o original. Foram identificadas na tradução apenas 17 ocorrências de *esserci*, enquanto o original possui quase três vezes mais ocorrências de "haver" com o sentido de existir (45 ocorrências).

Uma das hipóteses para isso, está no fato de que, durante a análise dos textos alinhados, foram identificadas mudanças nos textos traduzidos resultantes de escolhas tradutórias para os verbos lexicais que realizam processos existenciais nos textos originais, como pode ser observado nos Quadros 2 e 3, a seguir:

QUADRO 2
Mudanças nos processos observadas em *Legami familiari*
Legenda: Os processos aparecem em **negrito**.

Tipo de processo	**Texto original** *Laços de Família*	Texto traduzido ***Legami familiari***	**Tipo de processo**
Existencial	"Não **havia** ninguém no Jardim"	"Il giardino **era** deserto."	Relacional
	"**Havia** certas cousas boas porque eram quase nauseantes"	"Certe cose **davano piacere** proprio perché quasi nauseanti".	Mental
	" [...] a penumbra atormentada que sempre **há** nos meus sonhos quando de noite atormentado durmo."	" […] la penombra tormentata che sempre **popola** i miei sogni quando, di notte, tormentato io dormo."	Material
	"Em cada pessoa forte **havia** a ausência de piedade pelo cego [...].	"In ogni persona si **percepiva** come un'assenza di pietà verso il cieco [...]"	Mental

O Quadro 2 acima ilustra três tipos de mudanças tradutórias observadas no texto traduzido de *Laços de Família*, em que verbos lexicais que, no texto original, realizam processos existenciais foram traduzidos como verbos lexicais que realizam processos relacionais, materiais e mentais em *Legami familiari*.

Dentre essas mudanças, as que mais chamam a atenção são as duas traduções que apresentam verbos lexicais realizadores de processos mentais. Isso porque, as mudanças tradutórias para verbos lexicais realizadores de processos relacionais e materiais poderiam ser explicadas pelo fato de os processos existenciais estarem localizados no limite entre os mundos físicos e das relações abstratas, ou seja, entre os processos materiais e relacionais, como mencionado anteriormente e ilustrado pela Figura 1 no início do artigo.

No entanto, o mesmo não é válido para os processos mentais, que pertencem ao mundo da consciência. Dessa forma, as mudanças observadas se constituem como um elemento a ser observado durante o restante da pesquisa, no sentido de verificar se mudanças semelhantes ocorreram na tradução dos outros verbos lexicais realizadores de processos existenciais e se é possível estabelecer um padrão entre elas.

QUADRO 3
Mudanças nos processos observadas em *L'ora della Stella*
Legenda: Os processos aparecem em **negrito**.

Tipo de processo	**Texto original** *A hora da Estrela*	Texto traduzido ***L'ora della Stella***	**Tipo de processo**
Existencial	"Enquanto eu tiver perguntas e não **houver** resposta continuarei a escrever."	"Finché avrò domande e non **avrò** risposte continuerò a scrivere."	Relacional
	"E **tem** mais!"	"E non **è** tutto!"	Relacional
	"Sei que **há** moças que vendem o corpo [...]"	"**So** di ragazze che vendono il corpo […]"	Não-realizado
	"Agora me lembrei de que **houve** um tempo em que para me esquentar o espírito eu rezava: o movimento é espírito."	"Mi ricordo adesso di un tempo quando, per scaldarmi lo spirito, io pregavo: il movimento è spirito."	Não-realizado

O Quadro 3, por sua vez, também trás duas mudanças para verbos lexicais que realizam processos relacionais na tradução de *A hora da Estrela*. No entanto, as mudanças que mais chamaram a atenção durante a análise foram as verificadas na tradução de duas orações mentais que possuíam uma oração existencial encaixada. Na tradução de ambas as orações a oração existencial encaixada não foi mantida, ou seja, o significado existencial presente nelas não foi realizado.

Sendo assim, esse tipo de construção e de mudança também se tornam fatores a serem levados em consideração no restante da

pesquisa, visando verificar tanto se existem outras ocorrências de orações mentais com orações existenciais encaixadas no texto original como se o mesmo tipo de mudança pode ser observado na tradução das mesmas.

4 Conclusões

Este trabalho apresentou os dados coletados até o momento, juntamente com as analises preliminares realizadas no âmbito da pesquisa sobre processos existenciais que tem como objetivo identificar e analisar a ocorrências de orações existenciais no corpus monolíngue CALIBRA, bem como em um corpus paralelo bilíngue (português-italiano), visando contribuir para a descrição da realização dos significados existenciais no português brasileiro e para as pesquisas no campo dos estudos da tradução.

Apesar de os dados apresentados serem preliminares, já é possível argumentar que o tipo de análise realizada revela-se útil para esse tipo de investigação, uma vez que possibilita fazer observações sobre a ocorrência das orações existenciais tanto em textos em português de diferentes processos sócio-semióticos, como em textos traduzidos do português para o italiano.

No âmbito das ocorrências verificadas nos subcorpus Recomendar e Relatar do Corpus CALIBRA, foi possível não só verificar quais são os verbos lexicais realizadores de processos existenciais mais frequentes (haver, ter, existir, acontecer e aparecer) como que o subcorpus Relatar apresenta uma maior ocorrência e variedade de verbos lexicais realizando a existência.

Já no âmbito das ocorrências verificadas nos originais e nas traduções de *A hora da estrela* e *Laços de Família* da autora Clarice Lispector, foi possível verificar que as traduções apresentam menos ocorrências de orações existências que os originais, sendo que a maior disparidade foi observada na tradução de *Laços de Família*.

Uma das hipóteses para isso está no fato de terem sido verificadas mudanças tradutórias em que os verbos lexicais realizadores de processos existenciais nos originais em português foram traduzidos para o italiano como verbos lexicais realizadores de processos relacionais, materiais e mentais. Além disso, foram identificadas mudanças em que o significado existencial no texto original não é realizado no texto traduzido.

Por fim, dentre os próximos passos da pesquisa estão a expansão da análise para os outros processos sócio-semióticos (capacitar, compartilhar, explicar, explorar, fazer e recriar), no caso do corpus monolíngue, visando verificar não apenas a frequência em que os verbos lexicais realizadores de processos existências ocorrem nos textos que compõem os mesmos, como também, se eles revelam outros tipos de verbos lexicais realizando significados existenciais para além dos já identificados.

Já no caso do corpus bilíngue, além de concluir a análise dos outros tipos de verbos lexicais realizadores de processos existenciais nos textos originais e suas respectivas traduções, seria interessante realizar o processo inverso, ou seja, analisar as ocorrências das orações existenciais em italiano, para verificar se no texto italiano são construídos significados existenciais para traduzir outros processos no português.

Referências

FIGUEREDO, G. *Introdução ao perfil metafuncional do português brasileiro*. Tese (Doutorado) - Programa de Pós-Graduação em Estudos Linguísticos, Universidade Federal de Minas Gerais, 2011. (Inédita).

HALLIDAY, M. A. K.; MATTHIESSEN, C. *An introduction to functional grammar*. 3. ed. London: Edward Arnold, 2004.

LISPECTOR, Clarice. *Laços de família*. 9. ed. Rio de Janeiro: Rocco, 1998. 135 p.

LISPECTOR, Clarice. *Legami familiari*. Trad. Adelina Aletti. 4. ed. Milano: Giangiacomo Feltrinelli Editore, 1999. 121 p.

LISPECTOR, C. *A hora da estrela*. Rio de Janeiro: Rocco, 1998. (1. Edição, 1977)

LISPECTOR, Clarice. *L'ora della Stella*. Trad. Adelina Aletti. Milano: Giangiacomo Feltrinelli Editore, 1989.

SCOTT, M. *WordSmith Tools*. Oxford: Oxford University Press, 2007.

A chavicidade na análise de estilo em tradução: um estudo baseado em corpora paralelos espanhol/português

Célia Magalhães[1]
Ariel Novodvorski[2]

RESUMO: Este trabalho está inserido nos Estudos da Tradução baseados em Corpus (ETBC) e, em especial, nos estudos sobre Estilo em Tradução, entendido como atributo textual, na análise de um corpus paralelo de textos literários, no par linguístico espanhol/português. O estudo aborda a identificação das temáticas do corpus, por meio da chavicidade, entendida como qualidade textual (SCOTT, 2010; STUBBS, 2010; BERBER SARDINHA, 2009). O corpus de pesquisa está formado por três obras do autor argentino Ernesto Sabato, traduzidas ao português brasileiro por Sérgio Molina, e forma parte do Corpus ESTRA – Estilo em Tradução, desenvolvido no âmbito do LETRA/FALE/UFMG. Os procedimentos metodológicos usam subsídios da Linguística de Corpus, especialmente a utilização de ferramentas e utilitários do programa para análise lexical *WordSimth Tools*, 5.0 (SCOTT, 2008). Constatou-se a identificação de três campos semânticos, por meio da análise das palavras-chave, que indicaram a temática existencialista do corpus. A partir das semelhanças e diferenças observadas, constataram-se mudanças no ponto de vista narrativo, que afetariam o estilo e a representação mental dos leitores nos textos traduzidos.

PALAVRAS-CHAVE: Estudos da Tradução Baseados em Corpus, Estilo em Tradução, Chavicidade, Temática, Corpora Paralelos Espanhol/Português..

[1] Doutor em Literatura Comparada e Prof. Titular em Estudos Linguísticos: Estudos da Tradução na UFMG. Contato: celiamag@gmail.com.

[2] Mestre em Linguística Aplicada (UFMG). Professor do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU). Contato: ariel_novodovorski@yahoo.com.br.

ABSTRACT: This paper is affiliated to Corpus-based Translation Studies (CBTS) and to studies of Style in Translation, understood as a textual attribute. A parallel corpus of literary texts in the language pair Spanish/Portuguese is analysed and the study approaches text's aboutness through keyness, taken as a textual quality (SCOTT, 2010; STUBBS, 2010; BERBER SARDINHA, 2009). The research corpus consists of three works of Argentine author Ernesto Sabato, translated into Brazilian Portuguese by Sergio Molina, and it is part of ESTRA – Corpus for the Study of Style in Translation, designed by researchers at the Laboratory of Experimentation in Translation – LETRA, at UFMG. The methodological procedures come from Corpus Linguistics, especially the use of tools and utilities of the program for lexical analysis WordSmith Tools 5.0 (SCOTT, 2008). Three semantic fields were identified through keywords analysis, which indicated existentialism is one of the corpus themes or "aboutness". The investigation also points to shifts in the narrative point of view that would affect the style and mental representation of the readers of translated texts.

KEYWORDS: Corpus-Based Translation Studies, Style in Translation, Keyness, Text Aboutness, Spanish/BP parallel corpus.

1 Introdução

O rápido crescimento das pesquisas baseadas em corpus, desde a década de 1990, vem influenciado de maneira significativa os conceitos, estudos e ensino da tradução (LAVIOSA, 2002). Desse modo, uma aproximação ao campo dos Estudos da Tradução Baseados em Corpus (ETBC) torna-se necessária, no sentido de identificar as principais características e tendências já constituídas, com o propósito tanto de dar continuidade a uma tradição de pesquisa, como de contribuir para a expansão da área.

Pesquisadoras como Saldanha (2011), Malmkjaer (2003; 2004) e Baker (2000), entre outros, investigam aspectos de estilo em tradução, apoiadas nos subsídios advindos da Linguística de Corpus. Num esforço para identificar e mapear padrões que possam constituir marcas do tradutor nos textos traduzidos, Munday (2008) inclui o estudo de aspectos da ideologia, na análise de questões relacionadas ao estilo em tradução. Nesse sentido, as ferramentas da Linguística de Corpus são utilizadas, entre outros, para a investigação de aspectos

indicativos da presença da voz do tradutor e do levantamento de palavras-chave, por meio das quais é possível identificar a(s) temática(s) de um corpus de pesquisa e sua vinculação ao estilo dos textos.

Considerando que os ETBC se configuram como uma área interdisciplinar, recorreu-se aos estudos em estilística, especificamente àqueles que observam aspectos vinculados à análise da chavicidade, atrelada à identificação das temáticas do corpus (SCOTT, 2010; STUBBS, 2010; BERBER SARDINHA, 2009; SCOTT, 1998). Segundo Scott (2010), a análise da chavicidade possibilita, entre outras coisas, a identificação da(s) temática(s) do corpus e de indícios de estilo, uma vez que as palavras-chave funcionam como "ponteiros" que indicam áreas no corpus que seriam de interesse para o pesquisador.

Entre os objetivos específicos deste trabalho, destacamos (1) a intenção de identificar padrões recorrentes próprios dos textos traduzidos (TTs), com o intuito de constatar a noção de estilo em tradução enquanto atributo textual; e (2) o propósito de identificar a(s) temática(s) do corpus e possíveis mudanças que poderiam afetar o estilo nos TTs, em relação aos textos originais (TOs). As questões que buscamos responder neste estudo são: Haveria regularidades consistentes e significativas, na relação entre os TOs e TTs analisados, que confirmassem a noção de estilo em tradução? Quais seriam as temáticas presentes no corpus, a partir da análise das palavras-chave? Quais seriam as semelhanças e diferenças entre TOs/TTs, se comparadas as listas de palavras-chave?

O presente trabalho[3] traz uma contribuição para a constituição do Corpus ESTRA, na medida em que passa a expandir o material textual já existente, inaugurando uma linha de investigação. Esta nova perspectiva de análise assumida incorpora um subcorpus que compila traduções de obras literárias de um mesmo autor (Ernesto Sabato), feitas por um mesmo tradutor (Sérgio Molina) para o português contemporâneo, além de considerar em termos contrastivos o par linguístico espanhol/português, com pouca representação nas pesquisas anteriores. O principal objetivo na compilação do Corpus ESTRA é a investigação de traços de estilo na relação entre TOs e TTs, com a possibilidade de se mapear o estilo do tradutor,

[3] Pesquisa desenvolvida com o financiamento do CNPq, projeto 302178/2010-4, e Fapemig, PPM 0020/10.

verificando-se em que medida as mudanças na tradução podem afetar o estilo dos textos traduzidos. O Corpus ESTRA, até o momento da escrita do presente trabalho, possui um tamanho de em torno de seis milhões de itens (*tokens*), e está integrado, sobretudo, por traduções de textos originais feitas por diferentes tradutores, no par linguístico inglês/português.

Este artigo está composto por: (1) uma seção teórica, em que são apresentados os fundamentos dos ETBC, noções de Estilo em Tradução e os princípios básicos da pesquisa em chavicidade; (2) uma seção para a apresentação do corpus e dos procedimentos metodológicos; e (3) uma seção de análise e apresentação dos resultados. Por último, tecemos algumas considerações decorrentes deste estudo.

A próxima seção introduz o marco teórico em que se insere este trabalho.

2 Fundamentação teórica

2.1 Estilo em Tradução e os Estudos da Tradução Baseados em Corpus

Os *Estudos da Tradução baseados em Corpus* (ETBC) devem sua origem e inspiração tanto à *Linguística de Corpus* (LC) como aos *Estudos Descritivos da Tradução* (EDT). Segundo Laviosa (2002), ainda que as primeiras influências da LC nos estudos da tradução tenham surgido em função da busca pela constituição de uma metodologia de pesquisa mais coerente e efetiva, há elementos suficientes para assinalar que os ETBC passam a se configurar como um "novo paradigma". Para além de uma metodologia inovadora, a autora descreve as contribuições decorrentes de pesquisas desenvolvidas na área, entre outras, a formulação de hipóteses e construtos teóricos, a definição de ferramentas para análises empíricas, além das diversas aplicações.

Segundo Baker (2000), a estilística literária tem se concentrado, tradicionalmente, nas escolhas linguísticas conscientes por parte do escritor, justamente porque os estilistas literários estariam interessados nas relações entre as características linguísticas e a função artística, no modo como um escritor consegue determinados efeitos artísticos. Por outro lado, a estilística forense se concentra nos hábitos

linguísticos mais discretos, que vão além do controle consciente do escritor, e que frequentemente são registrados de modo subliminar pela audiência. Nesse sentido, Baker observa que, tanto nos estudos linguísticos como nos literários, a noção de estilo é associada (1) ao estilo de um escritor ou falante individual, (2) às características linguísticas associadas com os textos produzidos por grupos específicos de usuários da linguagem e em cenários institucionais específicos, ou (3) a características estilísticas específicas de textos produzidos num período histórico particular.

Baker (2000, p. 245) define estilo como "um tipo de impressão digital que é expressa numa variedade de características tanto linguísticas como não-linguísticas".[4] A autora ainda destaca ser muito comum, na tradução literária, que haja uma afinidade particular do tradutor com o escritor, o que possibilita a análise em função das escolhas. Sobre o estudo do estilo do tradutor, Baker assinala que deveria se concentrar no modo de expressão típica de um tradutor, e não simplesmente nas instâncias de intervenção explícita como, por exemplo, nas Notas do Tradutor e outros elementos paratextuais. Baker diz que os ETBC podem auxiliar na localização das expressões típicas que, para Hermans (1996), ficariam imperceptíveis.

Saldanha (2011), por sua vez, estabelece uma distinção entre estilo, entendido como um atributo textual, ou como um atributo pessoal. Com o propósito de definir o estilo do tradutor, ou seja, estilo enquanto atributo pessoal, a autora adapta a definição sobre a autoria de escrita formulada por Short (1996). Desse modo, Saldanha entende estilo como um modo de traduzir, reconhecível a partir de uma série de traduções de um mesmo tradutor, que diferencia o trabalho de um tradutor dos demais tradutores, por meio de padrões consistentes e distintivos de escolha. Esses padrões passam a ser característicos de um estilo pessoal e são independentes da referência ao estilo do autor ou do texto fonte.

Munday (2008) vai discutir estilo no contexto das marcas ou elementos linguísticos que tornam identificável um texto traduzido, ou série de textos, como o trabalho de um indivíduo em particular. O principal interesse do autor na investigação do estilo em tradução é verificar até que ponto o modo como padrões repetidos seriam

[4] No original: "a kind of thumb-print that is expressed in a range of linguistic - as well as non-linguistic - features".

representativos do estilo de tradutores individuais e como esses padrões poderiam afetar também a voz narrativa geral do autor do texto fonte, que ecoa através da voz do tradutor. Desse modo, Munday (*idem*, p. 19) vincula os conceitos de *voz* e de *estilo*, indicando que utiliza *voz* em referência ao conceito abstrato da presença do autor, do narrador ou do tradutor; e *estilo* em referência à manifestação linguística dessa presença no texto. O estilo do autor ou do tradutor somente poderia ser identificado pelo estudo da linguagem do texto, o que, em consequência, possibilitaria determinar a(s) voz(es) presente(s) no discurso, isto é, o estudo da *voz* deveria ser abordado por meio da análise de *estilo*.

Malmkjaer (2004, p. 14) observa que "*estilo* pode ser definido como uma regularidade consistente e estatisticamente significante de ocorrência no texto de certos itens e estruturas, ou tipos de itens e estruturas, dentre aqueles que são oferecidos pela língua como um todo".[5] Estilística tradutória é a metodologia de análise proposta por Malmkjaer (2004) e consiste no estudo de padrões recorrentes nas relações entre o texto fonte e a tradução, ou entre a obra mais vasta de um autor e um conjunto de traduções dessa obra. Na estilística tradutória se busca responder à questão sobre o porquê de uma tradução haver sido feita para significar de um modo determinado. Já na análise estilística monolíngue, segundo a autora, é possível fazer afirmações sobre o modo como qualquer texto provoca respostas na mente leitora, independente do modo como o texto tenha surgido. Malmkjaer define claramente seu foco de atenção no estilo dos textos traduzidos.

Boase-Beier (2006) vai considerar o estilo do texto original a partir da percepção do tradutor, sendo de seu interesse tanto o modo como é transmitido, modificado ou até conservado o estilo na tradução. A autora adota uma perspectiva cognitiva da estilística, entendendo estilo como uma expressão de estados mentais e visões de mundo. Segundo Saldanha (2011), há uma diferença fundamental, se comparadas as abordagens e metodologias de Malmkjaer e de Boase-Beier, por um lado, e a de Baker, por outro. Enquanto esta última se ocupa do estilo do tradutor, as primeiras abordam o estilo dos textos traduzidos.

[5] No original: "'Style' can be defined as a consistent and statistically significant regularity of occurrence in text of certain items and structures, or types of items and structures, among those offered by the language as a whole".

2.2 As *palavras-chave* na análise da temática e de estilo em tradução

Stubbs (2010, p. 21) assinala as características especiais que possuiriam as palavras-chave, destacando os significados sociais que expressam, por formarem parte do vocabulário de uma cultura e de uma sociedade, e em função do papel especial que desempenham nos textos, enquanto unidades de significado. Segundo o autor, nos trabalhos baseados em palavras-chave, não se deveriam separar as análises semântica e social.

As afirmações de Stubbs vão ao encontro da visão de Malmkjaer (2003; 2004), para a realização de toda análise estilística. A autora entende a necessidade de se considerarem fatores extralinguísticos imbricados, uma vez que funcionam como condicionantes da liberdade de expressão de um escritor, na hora de fazer seleções através das escolhas oferecidas por uma língua, num ponto particular de sua história. Desse modo, observa-se a necessidade de investigar as palavras-chave de cada um dos textos de Sabato que compõem o corpus de pesquisa e do conjunto das obras, com o intuito de alcançar os significados, incluindo aqueles a que se possam atribuir ideologias.

Stubbs analisa a expressão *palavras-chave* em três significados diferentes, a partir: (1) dos estudos culturais; (2) da análise comparativa e quantitativa de corpus, que identifica palavras estatisticamente proeminentes em textos ou coleções de texto; e (3) do trabalho em léxico-gramática. Para Stubbs (2010, p. 23), "Palavras-chave são tipos de icebergs: ponteiros para objetos lexicais complexos, que representam as crenças e valores compartilhados de uma cultura".[6] O autor conclui que o problema das palavras-chave é a grande diferença entre palavras individuais e o mundo social.

Conforme Scott (2010), embora ainda seja pouco compreendida, a análise da *chavicidade* (Keyness) está começando a despertar o interesse dos pesquisadores, como uma qualidade textual que daria fortes indícios sobre a temática do texto, junto a indicadores de estilo.

[6] No original: "Keywords are the tips of icebergs: pointers to complex lexical objects which represent the shared beliefs and values of a culture".

O autor ressalta que a chavicidade é uma qualidade intrínseca dos textos e não da linguagem em si; isto é, uma palavra é chave num dado texto ou corpus, e não numa dada linguagem.

Scott conclui que as palavras-chave funcionam como "ponteiros" para o pesquisador. A chavicidade, nesse sentido, indica áreas que valeria mais a pena serem investigadas, uma vez que essas palavras se tornam proeminentes por alguma razão que deveria ser analisada. O autor ainda assinala que a temática de um corpus não necessariamente será única, podendo haver diferentes temáticas num corpus.

Berber Sardinha (2009, p. 194) também aponta algumas das finalidades no uso das palavras-chave em análises linguísticas. Entre outros, o autor destaca a identificação da temática de um corpus ou de um texto, a descrição da organização interna dos textos, a localização de marcas indicativas de posicionamento ideológico e a possibilidade de traçar um perfil lexical de um autor ou de outros indivíduos.

Outra referência importante para o presente estudo é Scott (1998) que analisa um corpus paralelo com originais de Clarice Lispector traduzidos por Giovani Pontiero. Além da análise e identificação de traços de normalização, a pesquisadora fez o levantamento das palavras-chave e identificou, entre outros, os advérbios de negação "não" (TO) e "never" (TT). No contraste entre as listas de palavras-chave, Scott observou uma diminuição da negatividade na tradução, a partir da omissão de itens negativos ou de sua reformulação em itens positivos. Essa desconsideração, na tradução para o inglês, da recorrência de advérbios simples de negação do TO, pode haver afetado a recepção do TT na cultura de chegada, segundo Scott. Também são de importância as observações feitas pela pesquisadora a respeito da presença de dêiticos pessoais entre as palavras-chave.

A análise da chavicidade – entendida como qualidade textual – possibilita, entre outras coisas, a identificação da(s) temática(s) do corpus e de indícios de estilo, uma vez que as palavras-chave funcionam como "ponteiros" que indicam áreas que seriam de interesse para o pesquisador (SCOTT, 2010; STUBBS, 2010; BERBER SARDINHA, 2009; SCOTT, 1998). Também Aguiar (2010) observou a importância da análise das palavras-chave vinculadas ao campo semântico do corpo humano, para o estudo da recuperação da criatividade lexical nas traduções, vinculadas a supostas intenções semânticas e estilísticas do autor.

Nesse sentido e por meio da análise da chavicidade, uma das hipóteses deste trabalho é a identificação, por meio das palavras chaves, de campos semânticos que apontem para temática(s) do nosso corpus de estudo. Outra hipótese a ser verificada é se seria possível identificar marcas de estilo, a partir do contraste entre as listas de palavras-chave dos originais e das traduções.

A próxima seção apresenta o corpus de estudo e os procedimentos metodológicos desenvolvidos.

3 Corpus e metodologia

O corpus linguístico de análise, adotado para a realização desta pesquisa, é um subcorpus do ESTRA, um corpus compilado para o desenvolvimento de pesquisas interessadas na análise de Estilo em Tradução. Até o presente momento, o ESTRA possui em torno de seis milhões de palavras e passa por um tratamento que possibilitará o acesso online e a validação estatística.

A composição do corpus que será analisado neste trabalho reúne três obras literárias do escritor argentino Ernesto Sabato, escritas em língua espanhola, em sua variante rioplatense, e suas respectivas traduções feitas para o português brasileiro pelo tradutor literário Sérgio Molina. Os títulos dos TOs são *El túnel* (1982 [1948]), *Antes del fin: memorias* (1999 [1998]) e *La resistencia* (2000). As três traduções foram publicadas no mesmo ano (2008), pela editora Companhia das Letras, sob os nomes *O túnel* (2000), *Antes do fim: memórias* (2000) e *A resistência* (2008), sendo usada nesta pesquisa a 2ª reimpressão de *O túnel* e a 1ª reimpressão dos outros dois textos.

Tal como observado pelo próprio Sabato em cada uma das publicações, os textos correspondem a três tipologias diferentes de narrativas, a saber: ficção, memórias e epistolar, respectivamente. O Quadro 1 informa o nome das obras que compõem o Corpus de pesquisa, das editoras e datas de publicação.

QUADRO 1
Corpus de pesquisa

Obras	Autor/Tradutor	Editoras	Ano	1ª Publicação
El túnel	Ernesto Sabato	Sudamericana-Planeta	1982	1948
O túnel	Sérgio Molina	Companhia das Letras	2008	2000
Antes del fin: memorias	Ernesto Sabato	Seix Barral	1999	1998
Antes do fim: memórias	Sérgio Molina	Companhia das Letras	2008	2000
La resistencia	Ernesto Sabato	Seix Barral	2000	2000
A resistência	Sérgio Molina	Companhia das Letras	2008	2008

Os dados estatísticos do Corpus de pesquisa, apresentados a seguir na Tabela 1, foram obtidos com a função *Statistics* da ferramenta *Wordlist* do programa *WordSmith Tools*® (WST) em sua versão 5.0:

TABELA 1
Corpus de pesquisa

Textos	*Itens*	*Formas*	*Razão forma/item*
El túnel	31.741	5.183	16,33
O túnel	30.635	5.259	17,19
Antes del fin: memorias	31.379	7.099	22,68
Antes do fim: memórias	29.815	7.348	24,72
La resistencia	20.474	4.643	22,80
A resistência	19.599	4.811	24,70
TOTAIS	83.594 [es] 80.049 [pt] 163.643	12.120 [es] 12.568 [pt]	14,53 [es] 15,75 [pt]

Na leitura da tabela anterior, observamos que cada um dos TOs apresentou um número de *itens* superior a seu respectivo TT. Mas, em contrapartida, os TTs revelaram uma quantidade superior de *formas*, em relação a cada um dos TOs. Essa diferença registrou uma *Razão forma/item* superior nos TTs, fato que leva a pensar em uma maior diversidade lexical nas traduções. Berber Sardinha (2004, p. 94) aponta que "na prática, a razão forma/item indica a riqueza lexical do texto".

O resultado acima não confirma a hipótese já formulada em diversos trabalhos realizados com os corpora comparáveis,[7] no âmbito dos ETBC, de uma linguagem possivelmente menos variada na tradução, configurando-se como uma característica dos textos traduzidos se comparados a textos não traduzidos. Entretanto, confirmam resultados de trabalhos já realizados no LETRA, com corpora paralelos no par linguístico inglês/português. Por meio do levantamento, análise e comparação das listas de palavras-chave, assim como também pelo cotejo estrito das palavras lexicais e das gramaticais entre os subcorpora, em etapas futuras, seria mais um ponto de investigação a verificação de motivos para uma maior variedade lexical nos TTs do presente corpus em que as línguas são o espanhol e o português e em que se esperaria encontrar uma variedade lexical mais estabilizada entre os textos, dada a proximidade das línguas.

4 Procedimentos metodológicos

A seguir, apresentamos os passos metodológicos adotados para a realização do presente estudo:

1. Compilação e preparação do corpus de pesquisa: escaneamento (digitalização), aplicação do OCR, revisão e alinhamento, inserção de cabeçalho;
2. Leitura com a ferramenta *WordList*, para o levantamento dos dados estatísticos mais gerais do corpus e de cada um dos TOs e TTs;
3. Compilação do corpus de referência;
4. Levantamento e análise das palavras-chave, por meio da ferramenta *KeyWords*, e estabelecimento de campos semânticos;
5. Elaboração de quadros e tabelas e alinhamento dos textos para análise das palavras-chave.

[7] Baker (1995, p. 234) aplica o termo *corpora comparáveis* a duas coleções separadas de textos na mesma língua: uma, com textos originalmente produzidos em uma língua, e outra, com textos traduzidos para essa mesma língua. O resultado é um corpus monolíngue formado de textos não-traduzidos e textos traduzidos.

4.1 Compilação do Corpus de referência e levantamento das palavras-chave

Considerando as diferenças notadas ao comparar a *Razão forma/item* dos TOs e dos TTs entre si, o levantamento das palavras-chave foi considerado como um procedimento oportuno, uma vez que possibilitaria a observação contrastiva de aspectos linguísticos como a criatividade lexical e aspectos morfossintáticos, mas, principalmente, permitiria a identificação dos assuntos sobre os quais versam as obras, isto é, a temática do corpus, e se haveria diferenças nesse plano entre os TOs e TTs. Por meio da ferramenta *KeyWords* do WST, foi possível obter uma lista com as palavras-chave de cada um dos TOs e TTs, individualmente ou agrupados por línguas, considerando o Corpus geral de pesquisa.

Para efetivar o levantamento das palavras-chave, foi necessário compilar um Corpus de referência. A ferramenta *KeyWords* utiliza o Corpus de referência como ponto de comparação para poder determinar as palavras-chave do Corpus de estudo. De acordo com Berber Sardinha (2004, p. 100-102; 2009, p. 198), um Corpus de referência deve estar composto, no mínimo, por um número de *itens* 5 (cinco) vezes maior ao Corpus de estudo. Além desse detalhe, o autor destaca a importância de o Corpus de referência não conter o Corpus de estudo, e ainda explica que, como escolha não marcada para os estudos de palavras-chave, o Corpus de referência precisa possuir tipologias textuais diferentes daquelas do Corpus de estudo. A explicação dada pelo autor, sobre a importância de observar esses cuidados, consiste em que elementos característicos do Corpus de estudo poderiam ser filtrados pelo Corpus de referência, acarretando a perda de traços linguísticos que, do contrário, seriam considerados pelo programa como palavras-chave.

Uma vez adotados os critérios expostos acima, procedeu-se à compilação do Corpus de referência, com o intuito de levantar as palavras-chave. Para analisar paralelamente TOs e TTs, foi necessário compilar o Corpus de referência em dois subcorpora: um em espanhol e outro em português. Certificamos que a extensão em número de *itens* e a composição em termos de tipologias textuais diferentes, em ambos os subcorpora, guardassem o máximo de equivalência. Nesse sentido, foram compilados textos jornalísticos, acadêmicos e literários, via Internet, em proporções equivalentes, em língua espanhola e portuguesa.

Os textos que compõem o Corpus de referência foram salvos em formato *txt*, em duas subpastas, uma para o subcorpus em Espanhol e a outra para os textos em Português, e ainda separados em outras subpastas conforme fossem textos acadêmicos, jornalísticos ou literários, tudo dentro de uma pasta denominada CorREF01.

Com os textos em *txt* do CorREF01, foi possível levantar os dados estatísticos dos subcorpora, com a ferramenta *WordList*, a fim de assegurar-se que a extensão do Corpus de referência fosse, no mínimo, 5 vezes maior que o Corpus geral, e que houvesse um equilíbrio aproximado tanto na extensão quanto na diversidade de tipologias textuais entre os subcorpora. A Tabela 02 apresenta o número de textos, itens, formas e totais do CorREF01, para a língua espanhola e portuguesa.

TABELA 2
CorREF01

	Corpus de Referência Espanhol			***Corpus de Referência Português***		
	Textos	*Itens*	*Formas*	*Textos*	*Itens*	*Formas*
ACADÊMICO	9	171.975	15.341	11	175.784	19.911
JORNALÍSTICO	146	123.916	20.060	194	161.351	21.207
LITERÁRIO	8	209.368	21.952	6	168.765	20.150
TOTAIS	163	505.259	40.122	211	505.900	43.115

A ferramenta *KeyWords* compara as listas de palavras do Corpus de análise com as listas de palavras do Corpus de referência. Nesse sentido, foram feitas diversas listas de palavras-chave, com o intuito de contrastar os dados de cada um dos TOs e dos TTs separadamente. Cada uma das listas de palavras-chave foi salva no formato *kws* (*KeyWords*), para uma posterior análise contrastiva. Foram elaborados quadros e tabelas para disposição dos exemplos e dos dados coletados. Além desses procedimentos, foram observados alguns aspectos de colocação, tais como a co-ocorrência de itens individuais e agrupamentos lexicais (*clusters*), em torno de algumas das palavras-chave identificadas.

Para a obtenção destas primeiras listas de palavras-chave, utilizou-se a fórmula estatística *log-likelihood* e o valor de estatística p = 0,000001, que é o *default* do programa. Após esse procedimento inicial, o primeiro resultado foi de 210 palavras-chave em espanhol, das quais 30 foram palavras-chave negativas, e 199 em português,

das quais 29 foram negativas. De posse dessas listas, uma vez salvas em formato KWS, procedemos à organização dos termos, no sentido de facilitar as análises contrastivas.

Para isso, partindo desses dados, o passo seguinte foi proceder à separação das palavras em verbos, substantivos, adjetivos, advérbios, etc., para uma posterior identificação de campos semânticos de referência. Conforme cada uma das categorias léxico-gramaticais, o procedimento para organizar as listas de palavras-chave consistiu na separação dos termos, conforme as categorias de interesse. Desse modo, foi possível criar e salvar listas de palavras-chave apenas com substantivos, verbos, adjetivos, etc., tanto na língua espanhola como portuguesa, para análise posterior. A próxima seção apresenta a análise e a discussão dos resultados.

5 Análise e resultados

Com o intuito de identificar as temáticas do corpus de análise e, principalmente, no sentido de observar se haveria mudanças nesse plano de significação, mediante a comparação dos subcorpora dos TOs e TTs e com o auxílio da ferramenta *KeyWords*, procedeu-se ao levantamento das palavras-chave, entendendo que esse recurso subsidiaria também a detecção de mudanças de aspectos de estilo nos TTs. Desse modo, tentamos responder às seguintes perguntas: Quais seriam as temáticas presentes no corpus? Quais seriam as semelhanças e diferenças entre TOs/TTs, se comparadas as listas de palavras-chave? Em função de possíveis diferenças, quais poderiam ser as motivações para a proeminência de determinados itens em detrimento de outros? Que implicações poderiam ser observadas, na identificação de mudanças de estilo nos TTs?

As Figuras 1 e 2 ilustram de modo contrastivo as palavras lexicais, especificamente substantivos, classificados conforme a frequência, na primeira imagem, e a chavicidade, na segunda. Essas listas foram obtidas após uma organização do primeiro resultado, a partir da aplicação da ferramenta *KeyWords*, com o intuito de deixá-las, num primeiro momento, unicamente com os substantivos-chave e, desse modo, possibilitar a identificação das temáticas do corpus. Posteriormente, serão analisadas as demais classes de palavras-chave. Nas figuras, o destaque nos termos *tiempo/tempo*[8] deriva do interesse

[8] Por motivos de espaço, não apresentamos neste trabalho a análise decorrente do contraste desses termos.

em verificar a relação que poderia haver com questões de existencialismo (introspecção, distância, dúvida, etc.), além de possíveis diferenças em torno desse par, na consideração dos colocados e agrupamentos de palavras que co-ocorreriam com eles. Na Figura 02, a diferença de posição entre as palavras *tiempo* (20ª) e *tempo* (10ª), em termos de chavicidade, está justificada por um número maior de ocorrências nos TTs, na relação de 193 para 165 nos TOs.

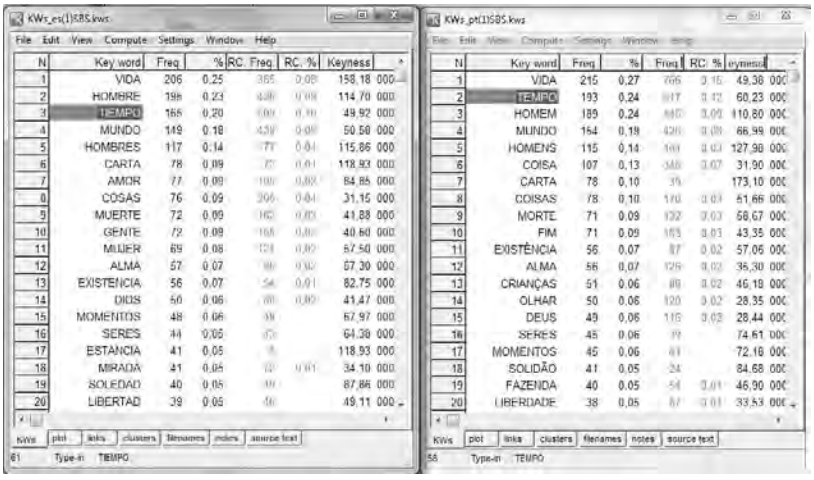


FIGURA 1 - Palavras-chave (substantivos) conforme a *Frequência*

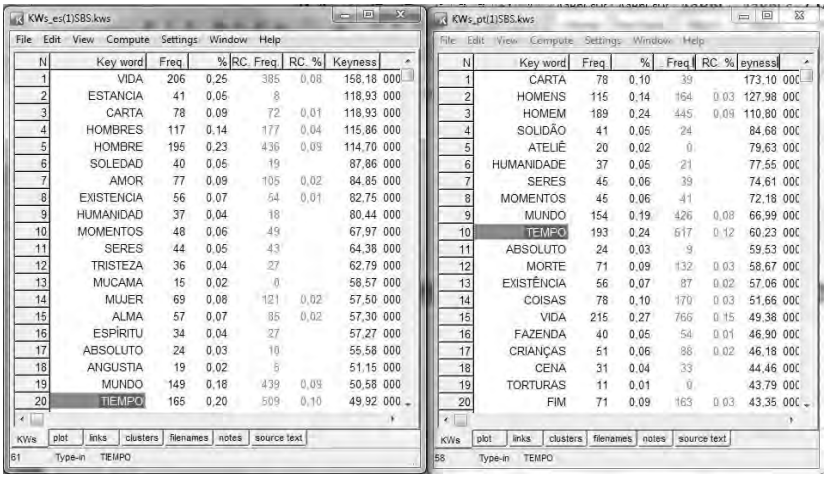


FIGURA 2 - Palavras-chave (substantivos) conforme a *Chavicidade*

Das 210 e 199 palavras-chave levantadas por meio da ferramenta *KeyWords*, respectivamente nos TOs e nos TTs, foram observados 61 substantivos no corpus em espanhol e 58 no corpus em português. Considerando essas listas de substantivos, os termos foram reunidos e classificados conforme a chavicidade, na tentativa de identificar a(s) temática(s) dos corpora e possíveis mudanças no contraste entre TOs/TTs.

A partir da leitura das listas de substantivos-chave e considerando ambas as direções, foi possível observar algumas diferenças no contraste TOs/TTs, no sentido de palavras que resultaram chave numa língua e não na outra. Essas listas também possibilitam uma aproximação, com auxílio da ferramenta *Concord*, a determinados termos que registraram maior ou menor frequência ou chavicidade nas traduções, propiciando uma análise comparativa mais específica.

Analisando as palavras-chave (substantivos) resultantes do contraste com o CorREF01, podemos apontar, a princípio, alguns campos semânticos em destaque e que guardam uma relação entre si, a saber: referências a aspectos *existenciais, seres, sentimentos/qualidades,* etc. O Quadro 02 agrupa os substantivos-chave que remetem a questões da *existência*, com suas frequências, tanto nos TOs como nos TTs.

QUADRO 2
Campo semântico - *Existência*

Substantivos-chave TOs	**Substantivos-chave TTs**
vida (206); *existencia* (56); *humanidad* (37); *alma* (57); *espíritu* (34); *mundo* (149); *tiempo* (165); *muerte* (72); *Dios* (50); *universo* (25); *eternidad* (14).	*vida* (215); *existência* (56); *humanidade* (37); *alma* (56); *mundo* (154); *tempo* (193); *morte* (71); *Deus* (49); *universo* (25); *eternidade* (13); *fim* (71); *destino* (34); *tempos* (35); *abismo* (15).

No quadro 2, observa-se que a palavra *espíritu* não resultou chave nos TTs, e as palavras *fim*, *destino*, *tempos* e *abismo* foram chave nos TTs, mas não nos TOs. O Quadro 03, a seguir, reúne os substantivos-chave que denotam *seres*.

O quadro 3 permite visualizar algumas diferenças entre os seres reportados como palavras-chave entre TOs/TTs. As palavras que denotam seres femininos, que resultaram chave nos originais, não constaram nas palavras-chave das traduções. Por outro lado, as expressões *crianças*, *operários*, *cego*, *anarquistas*, *multidões* e *mártires* não se mostraram chave nos TOs.

QUADRO 3
Campo semântico - *Seres*

Substantivos-chave TOs	**Substantivos-chave TTs**
hombres (117); *hombre* (195); *seres* (44); *mucama* (15); *mujer* (69); *gente* (72); *muchacha* (12); *pintor* (12).	*homens* (115); *homem* (189); *seres* (45); *crianças* (51); *operários* (12); *cego* (17); *anarquistas* (9); *multidões* (7); *mártires* (7).

Uma possível explicação para essas divergências entre as palavras-chave de uma e outra língua estaria na composição dos corpora de referência, uma vez que esses termos simplesmente não foram chave, mas constam nos corpora de estudo, com igual, maior ou menor frequência entre si.

O Quadro 04, a seguir, apresenta as palavras-chave que compõem o campo semântico de referência a sentimentos e qualidades.

QUADRO 4
Campo semântico – *Sentimentos/Qualidades*

Substantivos-chave TOs	**Substantivos-chave TTs**
soledad (40); *amor* (77); *tristeza* (36); *angustia* (19); *libertad* (39); *sentimientos* (28); *valores* (36); *desesperación* (19); *vanidad* (13); *fe* (22); *atributos* (12); *amargura* (13); *torturas* (12); *tranquilidad* (12); *emoción* (16); *miseria* (22); crueldad (11); *sufrimiento* (13); *ansiedad* (15); *ternura* (15); *odio* (20); *belleza* (20); *deseo* (23).	*solidão* (41); *tristeza* (36); *liberdade* (38); *valores* (37); *fé* (22); *amargura* (13); *torturas* (11); *miséria* (20); *ternura* (15); *horror* (18); *caos* (12); *dúvidas* (17); *pesadelo* (10).

No quadro acima, o resultado de substantivos-chave denotativos de sentimentos ou qualidades reportou um número maior de palavras nos TOs. Pelo contraste dos termos entre si, identificam-se expressões que vão desde uma semântica circunscrita por *solidão*, *tristeza*, *miséria*, *horror*, *caos*, etc., a outra que denota *liberdade*, *fé*, *ternura*, etc. A partir dessa lista de substantivos-chave, em etapa futura da pesquisa, ainda pode ser identificada a prosódia semântica dessas expressões, por meio da análise da co-ocorrência de adjetivos, e estabelecer de modo contrastivo se haveria diferenças entre TOs e TTs.

Além dos três campos semânticos identificados acima, há também na lista de substantivos-chave – ainda que menos recorrentes – termos denotativos de *lugares* (*fazenda*, *ateliê*, *estúdio*) e outros que

denotam *contato* entre *pessoas* (*mirada*, *conversación*, *carta*, *teléfono*). A seguir, apresentamos as considerações finais do presente estudo.

6 Considerações finais

Este trabalho, inserido no âmbito dos ETBC e, mais especificamente, nos estudos sobre Estilo em Tradução, propiciou a identificação de, a princípio, três campos semânticos (*existencialismo*, *seres* e *sentimentos e qualidades*), que apontam para os temas do corpus de pesquisa. Assim, por meio do levantamento e análise das palavras-chave, foi possível determinar algumas das temáticas do corpus, responder o questionamento sobre quais seriam os temas e confirmar as hipóteses sobre a identificação da temática do corpus, por meio da análise da chavicidade, conforme Stubbs (2010), Scott (2010), Berber Sardinha (2009) e Scott (1998).

Ainda, por meio da comparação das listas de palavras-chave, foi possível observar semelhanças e também algumas diferenças nos TTs, a saber: as palavras *fim*, *destino*, *tempos* e *abismo* foram chave nos TTs, no campo semântico denominado *existencialismo*, mas não nos TOs; as expressões *crianças, operários, cego, anarquistas, multidões* e *mártires*, do campo semântico denominado *seres*, foram chave apenas nos TTs; e, no campo semântico denominado *qualidades e sentimentos*, o número de palavras-chave (substantivos) nos TTs foi marcadamente menor, em relação aos TOs, mas foram registrados os termos *horror, caos, dúvidas* e *pesadelo*, que não constaram nos TOs.

A partir desse contraste, pode-se esperar que haja uma intensificação da instância final da existência, em algumas palavras que foram chave apenas nos TTs (*fim, destino, abismo*), além da proeminência de aspectos como a dúvida e o caos, e da focalização em alguns seres que não foram chave nos TOs (*crianças*, entre outros). Na consideração de alguns casos pontuais, espera-se, na etapa seguinte da pesquisa, verificar se o tradutor estaria motivado a explicitar, por meio da expansão de itens lexicais, as temáticas do corpus, tornando-as mais acessíveis para seus leitores. Isto é, o tradutor, enquanto leitor dos TOs, buscaria aproximar seu leitor da interpretação que ele teria feito durante a leitura dos originais.

Referências

BAKER, M. Towards a methodology for investigating the style of a literary translator. *Target*, Amsterdam, v. 12, n. 2, p. 241-266, 2000.

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

BERBER SARDINHA, T. *Pesquisa em Linguística de Corpus com WordSmith Tools*. Campinas: Mercado das Letras, 2009.

BOASE-BEIER, J. Translation and style: a brief introduction. *Language and Literature*, SAGE Publications (London, Thousand Oaks, CA and New Delhi), v. 13 (1), p. 9-11, 2004a.

BOASE-BEIER, J. *Stylistic Approaches to Translation*. Manchester: St. Jerome, 2006.

LAVIOSA, S. *Corpus-Based Translation Studies*: Theory, Findings, Applications. Amsterdam/New York: Editions Rodopi, 2002.

MALMKJAER, K. What happened to God and the angels: an exercise in translational stylistics. *Target*, Amsterdam, v. 15, p. 37-58, 2003.

MALMKJAER, K. Translational stylistics: Dulcken's translations of Hans Christian Andersen. *Language and Literature*. SAGE publications (London, Thousand Oaks, CA and New Delhi), v. 13 (1), p. 13-24, 2004.

MUNDAY, J. *Style and Ideology in Translation*: Latin American Writing in English. New York: Routledge, 2008.

SALDANHA, G. Translator Style – Methodological considerations. *The Translator*, v. 17, n. 1, p. 25-50, 2011.

SCOTT, M. *WordSmith Tools*, version 5.0. Liverpool: Lexical Analysis Software, 2008.

SCOTT, M. Problems in investigating keyness, or clearing the undergrowth and marking out trails... In: BONDI, M.; SCOTT, M. (Ed.). *Keyness in Texts*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2010. p. 43-57.

SCOTT, M. N. *Normalisation and reader's expectations*: A Study of Literary Translation with Reference to Linspector's 'A hora da Estrela'. Thesis (Doctorate in Philosophy). University of Liverpool, 1998.

STUBBS, M. Three concepts of keywords. In: BONDI, M.; SCOTT, M. (Ed.). *Keyness in Texts*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2010. p. 21-42.

Corpus de estudo

SABATO, E. *El túnel*. 4. ed. Buenos Aires: Editorial Ariel-Seix Barral Argentina S. A., 1984 (1948).

SABATO, E. *O túnel*. Trad. Sérgio Molina. 2. reimpr. São Paulo: Companhia das Letras, 2008 (2000).

SABATO, E. *Antes del fin*: memorias. 6. ed. Buenos Aires: Compañía Editora Espasa Calpe Argentina S.A./Seix Barral, 1999 (1998).

SABATO, E. *Antes do fim*: memórias. Trad. Sérgio Molina. 1. reimpr. São Paulo: Companhia das Letras, 2008 (2000).

SABATO, E. *La resistencia*. 2. ed. Buenos Aires: Editorial Planeta Argentina S.A.I.C./Seix Barral, 2000.

SABATO, E. *A resistência*. Trad. Sérgio Molina. 1. reimpr. São Paulo: Companhia das Letras, 2008.

O uso de *chunks* formados pelo verbo *get* por aprendizes de inglês como L2

Gláucio Geraldo Moura Fernandes[1]

RESUMO: Os chunks têm ganhado considerável atenção e empenho de pesquisa no campo de ensino/aprendizagem de L2. Muitos pesquisadores advogam o uso dos chunks no desenvolvimento da proficiência de aprendizes de língua inglesa L2, baseando-se na idéia de que uma parte importante da aquisição de línguas é a habilidade de compreender e produzir frases lexicais como um todo não analisável, ou "chunk", e que esses chunks são percebidos por aprendizes como padrões da língua. Neste trabalho, buscamos observar os caminhos que levam à aquisição, compreensão, e produção de chunks em língua inglesa L2. Para isso, nos embasamos em teóricos que propõem que as línguas podem ser agrupadas em categorias diferentes e traçam uma distinção tipológica entre línguas verb-framed e satellite-framed, assim como aqueles que dialogam a respeito da hipótese da transferência de língua e a respeito do uso de chunks para a aquisição de L2. Esse trabalho busca investigar o uso de chunks com o verbo get na formação de sentenças em inglês por parte de aprendizes e a correlação do seu uso para o português. Nessa perspectiva, observaremos a construção desses chunks tanto em uma língua satellite-framed, quanto em uma língua verb-framed.

PALAVRAS-CHAVE: chunks, línguas Satellite-framed e Verb-framed, língua inglesa L2.

ABSTRACT: Researchers have given considerable attention to chunks in the studies of L2 learning/teaching. They admit that the use of chunks develops learners proficiency in English as a L2, based on the idea that an important part of the languages acquisition is the ability of understand and produce lexical phrases as an analyzable thing, or "chunk", and that these chunks are noticed by learners as the language pattern. In this paper, we observe the ways that lead to acquisition, comprehension, and chunk's production in English as

[1] Mestre em Estudos Linguísticos pela Universidade Federal de Minas Gerais; Professor de Língua Inglesa; glauciocalama@yahoo.com.br

L2. We were based on researchers that propose that languages can be grouped in different categories and make a typological distinction between verb-framed languages and satellite-framed languages, as well as researchers that talk about the language transfer hypotheses and the use of chunks in the L2 acquisition. This paper tries to investigate the use of chunks with the verb get in construction of sentences in English by learners and their use in Portuguese. In this perspective, we will observe the construction of these chunks both in a satellite-framed language and in a verb-framed language.

KEYWORDS: chunks, Satellite-framed and Verb-framed Languages, English L2.

1 Introdução

Neste trabalho abordaremos a aquisição e o uso de certas 'construções' que representam um todo lexical em língua Inglesa, 'construções' estas cujos núcleos são verbos que indicam movimento, especificamente aquelas formadas pelo verbo *get*. Essas construções, também conhecidas como *chunks*, servirão como objeto de nossa pesquisa, nos fazendo refletir a respeito da sua aprendizagem por estudantes de Inglês L2.[2]

Os *chunks,* como um tipo de regularidade da língua no nível da forma e do significado, têm ganhado considerável atenção nos últimos anos nos estudos linguísticos em geral e no campo de ensino/aprendizagem de línguas em particular. Segundo Chang e Bao (2008), muitos pesquisadores advogam o uso dos *chunks* no desenvolvimento da proficiência de aprendizes de L2. Para esses autores, tal fato se baseia na idéia de que uma parte importante da aquisição de línguas é a habilidade de compreender e produzir frases lexicais como um todo não analisável, ou seja, um "*chunk*", e que esses *chunks* são percebidos por aprendizes como padrões da língua tradicionalmente pensados como 'gramática'.[3]

[2] A sigla L2, que significa Segunda Língua, será utilizada no decorrer do trabalho com o significado de Língua Estrangeira e Segunda Língua indiferentemente.

[3] Nesse caso, os aprendizes salientam o *chunk* como uma unidade gramatical cujo significado é dado pelo todo e não pelo sentido individual de cada uma das partes que o compõe.

Carrol (2000 *apud* Chang e Bao, 2008) argumenta que uma parte de nossa memória denominada de *memória de curto prazo*, por ser muito limitada no tamanho, pode sustentar poucas unidades de informação. Contudo, podemos aumentar essa retenção de informações agrupando pedaços individuais de informação em unidades maiores, e é a esse processo que damos o nome de *chunking*. Perspectivas lexicais defendem que a língua é constituída por *chunks* significativos que, quando combinados, produzem textos coerentes. De acordo com Chang e Bao (2008), os *chunks* de uma língua podem ser compreendidos e produzidos com pouco esforço e pouco processamento.

Por esse viés, buscamos, em nosso trabalho, observar os caminhos que levam à aquisição, compreensão, e produção de *chunks* em língua inglesa. Para isso, revisaremos trabalhos anteriores que tratam essas mesmas construções, denominando-as 'verbos frasais'. Posteriormente, nos embasaremos em teóricos que tratam a diferença entre línguas orientadas para o enquadramento verbal (*verb-framed*) e línguas orientadas para o enquadramento via partículas (*satellite-framed*),[4] assim como aqueles que discutem a hipótese da transferência de língua, levando-se em consideração a Linguística de Corpus como aporte metodológico. A partir da leitura, faremos uma coleta de dados com o propósito de construirmos uma análise de *chunks* com o verbo *get*, dados estes que serão analisados à luz de parâmetros explicitados no trabalho, procurando responder às questões que norteiam a presente pesquisa.

2 Justificativa

O interesse por este trabalho surge da minha experiência como professor de inglês L2. A questão relacionada aos *chunks* muito tem me chamado a atenção, principalmente com relação à forma através da qual estruturas são adquiridas por aprendizes do inglês L2.

A aquisição de *chunks*, na aprendizagem do inglês L1,[5] é considerada de fácil compreensão e produção por autores como Chang e Bao (2008). Por outro lado, para muitos outros autores, tal como Ellis (2003), é vista como um desafio para os aprendizes de Inglês L2.

[4] Os conceitos de línguas *verb-framed* e línguas *satellite-framed* serão explicados na sessão de Fundamentação Teórica. Tal qual SAMPAIO, SILVA E SINHA (2005), por convenção da literatura, manteremos estes termos em Inglês.

[5] A sigla L1 significa Primeira Língua e/ou Língua Materna.

Pesquisadores que trabalham numa perspectiva cognitiva, baseada no uso e na aquisição de línguas, como Ellis e Ferreira-Junior (2009), abordam padrões de frequência e aprendizado. Segundo eles, se a aprendizagem se dá a partir da exposição a construções particulares e do contato com uma determinada língua, então os aprendizes deveriam adquirir construções de alta frequência antes daquelas de menor frequência, e aqueles aprendizes de níveis mais avançados deveriam apresentar capacidade de construir estruturas a partir de fórmulas já memorizadas.

A proposta aqui apresentada, inspirando-se em pesquisas recentes a serem citadas na revisão bibliográfica, busca investigar o uso de *chunks*, mais precisamente os *chunks* com o verbo *get*, contrastando o seu papel no inglês L1 e o dos seus correspondentes em português L1, com fins à compreensão do seu uso no inglês L2 de falantes nativos de português L1. Nessa perspectiva, observaremos a construção desses *chunks* tanto em uma língua *satellite-framed* (no caso do inglês) quanto em uma língua *verb-framed* (no caso do português).

Como professor, venho percebendo que há uma grande dificuldade por parte dos estudantes de inglês L2 na construção de *chunks* com o verbo *get* e, buscando por trabalhos a respeito destas construções, cheguei à conclusão que não há estudos que abordem esse tópico em particular. Sendo assim, evidenciamos o caráter inovador da nossa pesquisa.

Esta pesquisa trará contribuições para a melhor compreensão do uso de *chunks* tanto no inglês, uma língua germânica, quanto no português, uma língua românica, assim como para a ampliação de formulações teóricas a respeito da aquisição de L2. Contribuições também surgirão no campo da Linguística de Corpus, já que os dados obtidos durante o trabalho serão analisados à luz de ferramentas e softwares disponíveis para esta tarefa.

A fim de se conhecer estas construções, sua aquisição e seu uso, esta pesquisa se justifica:

- Pela relevância em se averiguar como se dá a aquisição de tais expressões;
- Pela importância de se observar como os aprendizes usam tais expressões;
- Pela aplicabilidade do material resultante dessa pesquisa.

3 Objetivos e perguntas de pesquisa

O objetivo central desta pesquisa é identificar o uso de *chunks* com o verbo *get* em construções de movimento por aprendizes brasileiros de inglês L2. Esta tarefa se desdobra, necessariamente, em um estudo contrastivo entre três tipos de dados:

— corpus de inglês L1 como fonte de ocorrências de *chunks* produzidos por falantes nativos;

— corpus de português L1 como fonte de ocorrências de construções verbais de movimento produzidas por falantes nativos;

— corpus de inglês L2 como fonte de dados a serem analisados em cotejamento com os dois corpora anteriores com fins ao estabelecimento de potencial transferência linguística de L1 (português) para L2 (inglês).

Os objetivos específicos desta pesquisa são:

- Analisar as produções escritas dos participantes da pesquisa (aprendizes brasileiros) no intuito de se observar o uso dos *chunks* em inglês L2;
- Observar o comportamento dos *chunks* na língua inglesa L1 e seu contraste com os verbos de movimento encontrados na língua portuguesa L1, atentando para como os aprendizes de inglês L2 fazem uso dessas estruturas;
- Traçar hipóteses explicativas para os padrões encontrados no uso do inglês L2.

4 Fundamentação teórica

Considerando que este projeto busca investigar o uso de *chunks* formados pelo verbo *get* por aprendizes de inglês como L2, apresentaremos, primeiramente, uma abordagem mencionando trabalhos que tratam das construções 'verbos frasais', passando, logo após, a discorrer a respeito das línguas *satellite-framed* e *verb-framed*. Ao final, faremos uma abordagem referente à aprendizagem da L2 via *chunks*.

4.1 As Construções 'Verbos Frasais'

Linguistas com interesse de estudo em gramática (cf. VINCE, 1999, 2000; NELSON, 1996; BEAUMONT & GRANGER, 1992; MURPHY, 1997, 2004), os quais possuem uma abordagem focada na forma, tratam os 'verbos frasais' de modo a apresentar sua estrutura e explicitar o seu uso.

Segundo Beaumont & Granger (1992), 'verbos frasais' são verbos que mudam seus significados, de alguma forma, ao lhes serem acrescentadas partículas tais como *down*, *away*, *on*, *in*, *up*, *after*, *off*, *across*.

Há quatro tipos básicos de 'verbos frasais':

- O primeiro tipo é representado por aqueles que não tomam um objeto.

  Verbo + partícula

  Sit down
  Look out!

- O segundo tipo é composto por aqueles que tomam um objeto.

  Verbo + partícula + objeto

  I'll ***throw away*** the rubbish
  Take off your shoes.

  Verbo + objeto + partícula

  I'll ***throw*** the rubbish ***away***
  Take your shoes ***off***.

- O terceiro tipo é formado por aqueles que tomam um objeto, mas nos quais não se pode separar o verbo da partícula.

  Verbo + partícula + objeto

  Sue ***takes after*** her mother (Não: *Sue ***takes*** her mother ***after***)

- O quarto tipo é composto por três partes: um verbo + partícula + preposição como, por exemplo, *look after to*. Não podemos separar o verbo das outras partes.

  Verbo + partícula + preposição + objeto

  I'm ***looking forward to*** the weekend.

Já outros autores, tais como Nelson (1996), incluem à explicação precedente o fato de que, algumas vezes, podemos entender um 'verbo frasal' por meio de suas partes em separado, como no exemplo

*Please, **pick up** those papers*

onde ***pick*** significa *pegar* e ***up*** significa *para cima*, ou seja, *apanhar alguma coisa*, no caso da frase acima *aqueles papeis*.

Em outros casos, como

*They **put out** the fire*

o significado das partes em separado não é claro, pois ***put*** significa *colocar* e ***out*** significa *para fora*, ou seja, *colocar algo para fora*, no caso da frase acima *o fogo*. O significado do 'verbo frasal' nessa frase seria *fazer o fogo parar de queimar*.

Além disso, Nelson (1996) também acrescenta que alguns 'verbos frasais' têm mais de um significado, como os exemplos:

*The bomb **went off**. (It exploded)*
*The lights **went off** last night (They stopped working)*
*The milk **went off** last week (It went bad)*

Vince (2000, p.144) traz a seguinte explicação a respeito do uso dos 'verbos frasais': "O termo 'verbo frasal' é usado para designar os verbos seguidos por uma ou mais preposições, sendo também chamados de '*multi-word verbs*'. Os significados dos 'verbos frasais' não podem ser dados apenas pelo significado do verbo".[6]

A partir daí, o autor expõe o que já foi descrito por Beaumont & Granger (1992) acima. Há verbos com três partes, ou seja, aqueles 'verbos frasais' seguidos por um objeto. Vince (2000) também explica que os 'verbos frasais' são compostos por diferentes tipos. Segundo o gramático, estes verbos podem vir seguidos por um objeto ou o objeto pode vir entre o verbo e a preposição.

Murphy (1997, 2004) apresenta a mesma formação dos 'verbos frasais', definindo-os como verbos seguidos de palavras tais como ***in***, ***on***, ***up***, ***away***, ***round***, ***about***, ***over***, ***out***, ***off***, ***down***, ***back***, ***along***, ***forward***, etc., deixando claro que algumas vezes eles podem vir

[6] The term phrasal verb is used here for verbs followed by one or more prepositions. They are also called multi-word verbs. Their meaning cannot usually be guessed from the meaning of the verb on its own.

seguidos por preposições – *Why did you **run away** from me?* – assim como seguidos por objetos – *I **turned on** the lights*.

Autores com foco na abordagem comunicativa, tais como Celce-Murcia & Larsen-Freeman (1999) tratam as construções verbo + partícula, explicitando a partícula que segue o verbo como uma simples preposição ou um termo de fundamental importância para o entendimento da sentença, formando assim o 'verbo frasal'. Segundo estes autores, lidar com 'verbo frasal' é lidar com uma estrutura muito difícil para estudantes de inglês L2, pois o seu significado é frequentemente não-composicional, ou seja, o aprendiz pode saber o significado do verbo e o significado da partícula, mas quando colocados juntos, o significado único não é estabelecido. Outra hipótese para a dificuldade dos estudantes de inglês L2 aprenderem estas estruturas é que poucas línguas não-germânicas têm 'verbos frasais'. Vem daí os estudantes acharem tais estruturas estranhas e difíceis. Mas, apesar dessa dificuldade, é notório que ninguém pode falar ou entender Inglês, pelo menos no registro informal, sem o conhecimento dos 'verbos frasais'. Pelo fato de não perceberem isso é que alguns falantes não-nativos tendem a usar itens lexicais simples onde 'verbos frasais' seriam muito mais apropriados, como, por exemplo, em:

a. I ***arose*** early this morning.
b. I ***got up*** early this morning.

Apesar de a sentença a possuir um significado satisfatório, não é a forma apropriada numa conversação por não fazer parte do registro informal.

Outro desafio no aprendizado envolve condições que governam a separação opcional ou obrigatória entre os verbos e as partículas dos 'verbos frasais' usados transitivamente.

a. ***Turn out*** the lights.
b. ***Turn*** the lights ***out***.

Nesses dois exemplos a separação entre o verbo e a partícula é opcional, tendo em vista que o objeto direto não é um pronome.

Já nos exemplos

c. ***Turn*** them ***out***.
d. ****Turn out*** them.

a separação é obrigatória já que o objeto direto é um pronome.

Dessa forma, é possível observar que, na abordagem utilizada por Celce-Murcia & Larsen-Freeman (1999) são apresentadas razões de uso dos 'verbos frasais' e não regras.

Gramáticos que possuem um enfoque em corpora, ou seja, aqueles que buscam analisar as palavras extraídas de textos falados e/ou escritos, como Biber *et al* (2007, 2010), evidenciam que muitas unidades formadas por múltiplas palavras funcionam como verbos simples. Tal combinação de palavras tem, normalmente, um significado idiomático, ou seja, seu significado não pode ser dado por meio do significado de cada palavra em separado. O 'verbo frasal' é uma das classes dos verbos *multipalavras*.

Segundo Biber *et al* (2007, 2010), os 'verbos frasais' consistem em um verbo seguido de uma partícula adverbial (*verb + adverbial particle*), como por exemplo *carry out*, *find out*, *pick up*. Quando estas partículas adverbiais são usadas independentemente, elas têm sentidos literais que significam localização ou direção, como por exemplo, *out*, *in*, *up*, *down*, *on*, *off*.

Biber *et al* (2007, 2010) classificam duas subcategorias de 'verbos frasais': intransitiva e transitiva. Exemplos de 'verbos frasais' intransitivos são:

> **Come on**, tell me about Nick.
> **Hold on**! What are you doing there?
> I just **broke down** in tears when I saw the letter.

Já exemplos de 'verbos frasais' transitivos são:

> Did you **point out** the faults on it then?
> I ventured to **bring up** the subject of the future.
> I want to **find out** the relative sizes of the most common dinosaurs.

Os 'verbos frasais' transitivos permitem o movimento da partícula, ou seja, esta pode vir antes ou depois do objeto da frase. Como exemplo temos:

> I've got to **get** this one **back** for her mom.
> I went to Eddie's girl's house to **get back** my wool plaid shirt.

Já os 'verbos frasais' intransitivos possuem significados além dos significados individuais das palavras que os compõem. Temos como exemplo *come on, shut up, get up, break down, grow up, set in*, etc.

De acordo com Biber *et al* (2007, 2010), os 'verbos frasais' intransitivos mais comuns são aqueles que representam uma atividade, como *get on, look out, move in, step up, walk in.*

Com relação à produtividade dos verbos e das partículas adverbiais, os verbos que são mais produtivos em se combinar com partículas adverbiais para formar 'verbos frasais' estão entre os verbos lexicais mais comuns. Dentre eles, podemos citar os verbos *take, get, come, put, go*.

É a partir das abordagens a respeito dos 'verbos frasais' adotadas pelas gramáticas de referência que surgem os estudos baseados em *chunks*, principalmente no que se refere à produtividade de construções verbo + partícula.

Pelo fato de a abordagem dos estudos tradicionais possuir uma análise dos 'verbos frasais' limitada à forma, a opção teórica por nós adotada neste trabalho vai abordar a funcionalidade e uso dessas construções, construções estas que serão chamadas ao longo do trabalho de *chunks*, usando assim a terminologia proposta por Ellis (2003).

Sendo assim, nos utilizaremos dessas construções – *chunks* – buscando entender como a sua aquisição por aprendizes de Inglês L2 se dá, contribuindo, assim, para a expansão do escopo dos estudos tradicionais, os quais possuem uma abordagem primordialmente focada na forma.

4.2 Línguas *Verb-framed* X Línguas *Satellite-framed*

A organização lingüística de eventos de movimento pode ser realizada de maneiras diferentes, em línguas diferentes. No entanto, segundo Sampaio, Silva e Sinha (2005), a organização dos eventos de movimento pode ser descrita num conjunto bastante limitado que ampara padrões universais. Talmy (2000) propõe que as línguas podem ser agrupadas em duas categorias diferentes e traça uma distinção tipológica entre línguas *verb-framed* (VF) e *satellite-framed* (SF). As línguas VF são aquelas que expressam alguns significados associados à direção do movimento (*Path*) no verbo, como no português: *entrar, sair, subir, descer*. Observa-se que em línguas desse tipo a noção da direção do movimento está na própria estrutura verbal, sendo redundante se dizer *sair para fora*, porque *sair* já significa, por si só, movimento *para fora*. Já as línguas SF têm a tendência de expressar a direção do movimento por meio de uma frase

preposicional (PP), incluindo partículas associadas ao verbo, como no inglês: *go in, go out, go on, go down.* Observa-se que as línguas SF não apresentam flexão na estrutura do verbo na sua organização morfossintática, mas o verbo é obrigatoriamente seguido de uma partícula (satélite) que é o elemento léxico/gramatical que traz a carga semântica da direção do movimento.

Segundo Sampaio, Silva e Sinha (2005), uma língua VF dispõe de um sistema em que a direção do movimento está expressa no verbo principal da sentença, por outro lado o modo do movimento fica expresso em um verbo que não indica a direção, enquanto que a maneira (*manner*) como acontece o movimento, geralmente vem expressa por uma construção de gerúndio:

*O menino **saiu** da casa corre**ndo**.*

Nas línguas SF, o movimento e o modo estão no verbo principal, enquanto que a direção do movimento se dá numa partícula (satélite) associada ao verbo:

*The boy run **out** of the house* (*O menino correu **pra fora** da casa*)

Talmy (2000) endereça padrões tipológicos e princípios universais sublinhando a descrição dos eventos de movimento. Em particular, ele está interessado em questões às quais elementos semânticos, tais como movimento, direção, figura, terreno, maneira ou causa, são expressas por determinados elementos de superfície, tais como verbos, adposições, orações subordinadas ou satélites. Talmy (2000) assume que os eventos de movimento têm quatro partes constituintes, chamados de Figura, Terreno, Direção e Movimento. A Figura é um objeto que se move ou se encontra em um determinado lugar relacionado a outro objeto, o Terreno. O Movimento é caracterizado como a presença por si no evento do movimento ou localização. A Direção é o curso seguido ou espaço ocupado pela Figura com relação ao Terreno. Maneira e Causa são vistos como eventos externos distintos que podem ser configurados como co-eventos para um evento de Movimento. A relação que tais co-eventos podem manter com os eventos de Movimento podem ser múltiplos, mas os principais discutidos por Talmy (2000) são a Maneira e a Causa.

De acordo com Talmy (2000), satélites não são categorias sintáticas particulares, mas se encontram em uma relação gramatical particular com o verbo. São caracterizados como constituintes

imediatos da raiz do verbo, mas não são inflexões, auxiliares, ou argumentos nominais e assumem estarem relacionadas à raiz do verbo como elementos periféricos (ou modificadores) de um núcleo. Uma raiz verbal unida a este satélite, então, forma um constituinte, o "verbo complexo".

Com relação à aquisição de línguas SF, como o inglês, e a formação de predicados tidos como complexos, Snyder (2001) foca seu trabalho nas estruturas argumentais, ou seja, estruturas que são tipicamente analisadas como predicados complexos ou construções oracionais menores. Alguns exemplos em Inglês são os resultativos (nos quais o verbo principal combina com a frase adjetival (AP), resultando a uma ação) – *John painted the house red* – e as construções verbo-partícula (nos quais o verbo principal combina com uma partícula pós-verbal) – *Mary* ***picked up*** *the book* ou *Mary* ***picked*** *the book* ***up***. Em ambos os exemplos das construções verbo-partícula, um verbo principal combina com um predicado secundário para formar uma nova expressão que semanticamente se assemelha a um simples verbo.

Dado que as línguas românicas sistematicamente carecem de construções verbo-partícula e outras construções que são comumente analisadas como predicados complexos, Snyder (2001) conclui que predicados complexos do tipo encontrado no inglês são semanticamente excluídos destas línguas românicas, como é o caso do português.

Snyder (2001) mostra que as construções que envolvem a formação de predicados complexos, na sintaxe de uma dada língua que permite esse tipo de derivação, são adquiridas como um grupo. Ele propõe a disponibilidade morfológica da composição produtiva da raiz (verbo), ou seja, a disponibilidade marcada na sintaxe para a produção de complexos sejam eles adjetivais, preposicionais, nominais ou verbais, como um pré-requisito crucial para a emergência de predicados complexos tais como os *chunks* em análise neste trabalho. Tais composições se dão sem que haja mudança da classe morfológica do elemento raiz. Por isso tais compostos são nomeados como endocêntricos, uma vez que a raiz determina a classe do composto. Por exemplo, no inglês é possível termos formas como *tea cup*, onde *cup* é a raiz, que compondo-se com o modificador nominal *tea*, permanece inalterada em sua classe morfológica, i.e., permanece como um nome. Isto é expresso pelo parâmetro de composição.

> A gramática [não licencia, licencia] formação de compostos endocêntricos durante a derivação sintática. [valor não marcado] (Snyder, 2001, 328)

Tal relação apresenta a seguinte previsão. Primeiro, a disponibilidade de predicados complexos e a disponibilidade de raízes (verbos) produtivas deveriam ser modelos próximos para a composição de uma língua em particular. Segundo, na aquisição do inglês como primeira língua, por crianças, a idade na qual tais predicados complexos são usados produtivamente num primeiro momento deveria ter uma correspondência próxima à idade na qual a composição de raízes (verbos) é produzida. Snyder mostra que ambas previsões são relevantes.

Segundo Snyder (2001), línguas com um aspecto positivo de parâmetro de composição, como o Inglês, empregam produtivamente os compostos N-N e podem também criar novos compostos endocêntricos de raiz, como mencionado acima. A título de exemplo temos:

banana box

Línguas com um aspecto negativo de parâmetro de composição, como o português, não (pelo menos produtivamente) permitem tais composições. A título de exemplo temos

*caixa banana

Nesse caso, seria necessária uma construção exocêntrica, do tipo [N[Prep N]].

Snyder (2001) mostra que as construções que envolvem formação de predicados complexos em sintaxe são adquiridos como um grupo – *chunk* – nas línguas que permitem esse tipo de derivação. O autor afirma que as mesmas línguas que permitem tais tipos de composição empregam produtivamente construções verbo-partícula e resultativos adjetivais. Desse modo, Snyder disponibiliza evidências a respeito da aquisição do inglês, afirmando que estes tipos de formação de predicados complexos são adquiridos logo depois que os compostos endocêntricos de raiz são adquiridos.

Snyder (2001, p.19) aborda características semânticas que unificam as construções de predicados complexos, os quais derivam de uma possibilidade de composição semântica disponível em compostos endocêntricos. Nessa perspectiva, o autor, abordando a restrição no modo de composição semântica, propõe a Restrição do Predicado

Complexo nos mostrando que duas expressões sintaticamente independentes podem conjuntamente caracterizar o tipo de composto endocêntrico em um ponto de interpretação semântica.

De acordo com a explanação de Snyder (2001) a respeito da Restrição do Predicado Complexo, pode-se fazer a seguinte análise: na frase "*John **went out** of the class furiously*", tomando aqui a terminologia de Parsons (1990, apud Snyder, 2001, p.19), observa-se uma subparte *"development"* (a ação representada pelo verbo '*go*') e uma subparte *"culmination"* (o evento, caminho, descrito pela partícula '*out*'). Ambos, '*go*' e '*out*' participam da caracterização do tipo de evento descrito pela expressão. Nesse caso '*go*' contribui com o desenvolvimento e '*out*' contribui com a culminação, de uma realização de um tipo de evento.

Segundo Snyder, essa relação entre '*go*' e '*out*' só é possível se estas partículas forem subpartes de um composto endocêntrico em um ponto de interpretação semântica. Ainda, essas expressões funcionam claramente de forma independente na sintaxe, como evidenciado pelo fato que elas são descontínuas na estrutura da superfície da sentença. Dessa forma, a formação de compostos endocêntricos relevantes devem se estabelecer durante a derivação sintática, e tal formação de compostos na sintaxe é precisamente possível porque o inglês toma o aspecto tratado no "Parâmetro de Composição" citado anteriormente.

Snyder (2005) argumenta que as línguas com um aspecto positivo quanto ao parâmetro de composição são as mesmas línguas que são agrupadas como SF por Talmy, enquanto aquelas línguas com aspectos negativos coincidem com as línguas VF. Essa correlação permite estender o parâmetro de composição para construções envolvidas na descrição de eventos de Movimento. Snyder mostra que isto é sustentado pelos fatos da aquisição, uma vez que as primeiras combinações da Maneira dos verbos de movimento com frases preposicionais dentro de predicados complexos se correlacionam muito proximamente aos primeiros usos dos novos compostos N-N.

4.3 Aprendizagem da L2 via *chunks*

Pesquisadores interessados na aquisição de L2 têm estudado os processos que levam ao aprendizado; dentre estes grande ênfase tem sido dada aos processos que levam à transferência de L1 para L2. Segundo Ellis (2003, p.70), os aprendizes de inglês L1

desenvolvem padrões que são frequentemente baseados em *chunks* de palavras ou frases os quais contém espaços que podem ser preenchidos por uma variedade de palavras, por exemplo, subgrupos de substantivos ou verbos (*I don't + Verbo*; *I can't + Verbo*; *Where's + Substantivo + gone*?). Ellis (2003) propõe essa sequência na aquisição de L1 e, segundo Moraes Bezerra (2003), são modelos como esse, relacionados ao processo de aquisição de L1, que têm sido tomados em pesquisas na área de aquisição de L2.

Pensando na construção do conhecimento lingüístico de não-nativos, Selinker (1992) propõe o conceito de *Interlíngua*. Este autor apresenta alguns modelos teóricos sobre os estudos de aquisição/ aprendizagem de L2, modelos estes que têm enfocado a interlíngua durante as últimas décadas. Um modelo teórico tratado pelo mesmo autor e que fora proposto por Lado[7] (1957, *apud* SELINKER, 1992) é a chamada *transferência entre línguas*.

A transferência entre línguas (*language transfer*) e a influência que há entre línguas (*cross-linguistic influence*) são comumente empregadas em pesquisas em L2. Segundo Odlin (2003, p. 436), "a transferência é a influência que resulta nas similaridades e diferenças entre a língua alvo e qualquer outra língua que tenha sido previamente adquirida".[8]

De acordo com Ellis (2003), vale ressaltar que a aquisição de uma L2 é diferente da aquisição de uma L1 em vários aspectos, sendo um destes a transferência que pode ser feita da L1 para a L2. Nesse caso, adultos têm já adquirido conhecimento de categorias sintáticas abstratas e itens lexicais em sua L1, e esse conhecimento pode guiar a combinação na interlíngua com a L2 em vários aspectos, podendo estes serem bons (transferência positiva) ou ruins (transferência negativa).

A comparação da interlíngua com a L1 e a L2 tem certas limitações, especialmente com relação à transferência positiva. Segundo Odlin (2003), se a L1 e a L2 mostram pouca ou nenhuma diferença em alguma estrutura comum a ambas, qualquer padrão de transferência positiva não deveria diferir muito, e qualquer diferença que se encontre nos padrões de interlíngua, nestes casos,

[7] LADO, Robert. *Linguistics across cultures:* Applied linguistics for language teachers. Ann Arbor: Michigan University Press, 1957.

[8] "Transfer is the influence resulting from the similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired".

não dirá muito a respeito da transferência. Isso resulta no interesse por se entender outros fatores na aquisição de uma L2.

Segundo Ellis e Ferreira-Júnior (2009), outros fatores responsáveis por guiar a aquisição seriam a freqüência, a genericidade semântica, e a prototipicidade. Ellis e Ferreira-Júnior argumentam que fatores baseados no uso (freqüência) e escopo semântico (genericidade e prototipicidade) contam para a aquisição de construções verbais específicas por aprendizes.

Com relação à aquisição de verbos, Ellis e Ferreira-Júnior (2009) argumentam que os aprendizes, em um processo de aquisição naturalístico, aprendem primeiro os mais frequentes, prototípicos e genéricos. Segundo os autores, formas mais salientes são mais fáceis de serem adquiridas por aprendizes de inglês como L2, enquanto outras formas são ofuscadas e até bloqueadas dificultando a aquisição.

De acordo com Ellis (2003), para que uma sentença seja aprendida é necessário, por parte do aprendiz, uma repetição contínua. À medida que o aprendiz é exposto à língua, o desempenho é desenvolvido em diversos patamares: a proporção do léxico corretamente empregado, a produção acurada de sentenças e expressões, uso apropriado de bigramas e trigramas,[9] e a conformidade na probabilidade sequencial a nível de letras, palavras e frases. Segundo o autor, as sequências ortográficas, assim como os *chunks*, que são repetidos durante o processo de aprendizagem, são lembrados mais facilmente.

Segundo Newell (1990, *apud* Ellis, 2003, p.76),

> um *chunk* é uma unidade da organização da memória, constituído pela junção de um grupo de elementos já formados (os quais podem ser *chunks*) na memória e os unindo numa unidade maior. *Chunking* consiste na habilidade de construir tais estruturas repetidamente, conduzindo assim a uma organização hierárquica na memória. *Chunking* parece ser um aspecto comum da memória humana.[10]

[9] As letras podem co-ocorrer em bigramas, trigramas e outras regularidades ortográficas. (Ellis, 2003, p.75)

[10] A chunk is a unit of memory organization, formed by bringing together a set of already formed elements (which, themselves, may be chunks) in memory and welding them together into a larger unit. Chunking implies the ability to build up such structures recursively, thus leading to a hierarchical organization of memory. Chunking appears to be a ubiquitous feature of human memory.

Chunks formados por verbos, como o verbo *get* com o qual trabalharemos nessa pesquisa, são mais difíceis de serem aprendidos devido a possibilidades de várias entradas com determinados verbos. Segundo Ellis (2003) há construções que são mais abstratas como aquelas que indicam um movimento atribuído pelo *verbo + path* (trajetória). No exemplo *"If we get off early from work, we'll go fishing"* nós apenas saberemos o significado do *chunk* presente na frase se memorizarmos tal construção. De acordo com Ellis (2003), o aprendizado de construções abstratas é mais intrigante, iniciando com um *chunking* e uma fórmula que provém da memória.

O verbo *get*, por ser um verbo que necessita de um *path* (partícula de movimento) ligado a ele para que o sentido seja estabelecido, exige do aprendiz um maior conhecimento da língua inglesa. Dessa forma, fatores que influenciam neste processo, tido como fatores psicolingüísticos – relacionados à frequência de ocorrência dos elementos da língua, assim como a sensibilidade em se associar certos atributos a certas coisas através do uso que fazemos da língua –, conspiram na aquisição e no uso de qualquer construção lingüística, dentre estas as construções destes *chunks*. Como dito anteriormente, outro fator que contribui para esse aprendizado é a repetição devido ao fato de que os *chunks* ativam representações de significado que tornam a sequência mais saliente nas entradas com determinados verbos. Quando o aprendiz usar esses *chunks* novamente, eles tendem a salientá-lo como unidade.

Pode-se dizer que a freqüência promove o aprendizado e, segundo Ellis (2002a, p.146), a psicolingüística demonstra que os aprendizes de uma língua são requintadamente sensíveis às entradas em todos os níveis, isto é, aprendemos a classificar os elementos de nosso mundo sem sermos instruídos para tal, apenas pela freqüência com a qual ouvimos e vemos.

De acordo com alguns teóricos, é possível observar que aprendizes de todos os níveis de proficiência são vistos como memorizando *chunks* de alta frequência que contribuem para a formação das categorias de protótipos funcionais. Essa memorização de *chunks* pode servir de auxílio aos aprendizes na aquisição da L2, o que deve ser tomado com certo cuidado para que os significados de exemplares da L1 não sejam simplesmente *transferidos* para a L2, esquecendo-se da diferença existente entre as línguas. O hábito de comparação de palavras e expressões na língua inglesa pode levar o

aluno a interpretações confusas, induzindo-o a erros. Estes erros podem ser gráficos, fonológicos, gramaticais e/ou léxico-semânticos.

Sendo assim, torna-se relevante entender melhor o processo de aquisição, uso e compreensão de *chunks*, no caso de nosso trabalho *chunks* formados pelo verbo *get*, por aprendizes de Inglês como L2, além de compreender questões ligadas à natureza da linguagem, da aprendizagem humana e mesmo em relação à comunicação.

5 Metodologia

5.1 Participantes

Nosso trabalho tomará como sua base de dados produções linguísticas de 52 alunos/aprendizes de inglês como L2, estudantes do curso de Letras da Universidade Federal de Minas Gerais/UFMG e Universidade Federal de Ouro Preto/UFOP. Todos os participantes são falantes nativos do Português e têm esta como a língua falada no dia-a-dia como meio de comunicação. Estes aprendizes estão classificados em um nível intermediário avançado de aprendizado e têm, provavelmente, quatro anos de estudo da língua Inglesa em contexto de sala de aula.

Os dados foram obtidos através dos participantes e por meio de uma gama de atividades que buscaram estimulá-los a criar determinadas produções, por meio de insumos escritos, orais ou visuais, almejando o uso de construções – *chunks* – com o verbo *get*. Estas atividades incluíram descrição de figuras e observação de um filme e recontagem de sua história por escrito. A regra da língua portuguesa L1 e da língua inglesa L1 foi tomada para servir como um comparativo com a produção dos falantes de Inglês L2.

5.2 Procedimentos

A produção escrita dos dados produzidos pelos participantes foi recolhida por meio de duas sessões que consistiram, como dito anteriormente, na descrição de figuras e observação de um filme e recontagem da história por escrito. Enquanto nenhum procedimento possa infalivelmente eliciar as estruturas que desejamos que os participantes produzam, as atividades foram criadas e desenhadas para disponibilizar contextos que encorajassem os alunos na

construção de sentenças com *chunks* com verbos de movimento. Os verbos de movimento foram disponibilizados, de forma implícita ou explícita, em ambas as figuras e filme, de maneira que os participantes não criassem sentenças que fugissem do nosso propósito. Distratores não foram usados, de forma que todos os contextos descrevessem ou representassem cenas comumente expressas por verbos de movimento.

A primeira sessão para a compilação do corpus consistiu na tarefa de *fill in the blanks.* Foi entregue aos participantes 13 imagens as quais indicavam um movimento expresso pelos personagens apresentados. Juntamente com estas, apresentamos sentenças as quais os participantes deveriam completar usando o verbo que melhor se enquadrasse com relação à ação desempenhada pelo personagem presente na imagem.

A segunda sessão para a compilação do corpus consistiu na tarefa de observação de um pequeno filme, o qual expressa movimento, e recontagem da história pelos participantes. Nessa recontagem, observamos os possíveis usos de *chunks* formados pelo verbo *get* seguido por partículas indicando o caminho explicitado. O filme utilizado para esta tarefa foi o *The Pear Film*, um filme de seis minutos de duração desenvolvido por Wallace Chafe em 1975 na Universidade da California. Este filme está disponível em <http://www.linguistics.ucsb.edu /faculty/chafe/pearfilm.htm>.

Nessa perspectiva, observamos a capacidade de uso de *chunks* (produção linguística) formados pelo verbo *get + Path,* por aprendizes de inglês como L2, levando em consideração o fato de que o inglês é uma língua *Satellite-framed*, assim como explicitado por Talmy (2000).

Como corpora de comparação, utilizamos transcrições de narrativas desenvolvidas a partir da recontagem da *Pear Story* tanto por falantes de inglês L1, quanto por falantes de português L1. As narrativas dos falantes de inglês L1 foram obtidas por meio do site *"The Chinese Pear Story"*, disponível em <http://www. pearstories.org/english/english. htm>. Nesse site foi possível encontrar transcrições de 20 narrativas, um número comparável com o que será compilado em nosso trabalho.

Dado o fato de não encontrarmos corpora de falantes de português L1 comparáveis ao corpus a ser compilado nessa pesquisa, decidimos coletar as narrativas dos alunos/participantes também em português. Essa coleta se deu da mesma forma que foi compilado

o corpus de estudo. Assim, conseguimos o minicorpus das narrativas em três versões, o que nos possibilitou, a partir dos resultados encontrados, fazer uma análise comparativa das formas e suas frequências, além de termos acesso à produção em língua portuguesa L1 e cotejá-la com aquela em língua inglesa L2 em busca de marcas de transferência e interlíngua.

A natureza de nossa pesquisa é empírica com eliciação de dados a partir da exibição de figuras e de um filme que trazem em sua estrutura conceitual ações desempenhadas por construções verbais de movimento. Não objetivamos fazer nenhum procedimento experimental com intuitos de mensuração perceptual. Nosso interesse neste trabalho circunscreve-se a tarefas de produção.

A transcrição dos dados obtidos através da produção dos aprendizes foi analisada por meio do software TextSTAT. O TextSTAT é um programa muito simples para análise de textos, o qual possui a capacidade de ler textos completos (em diferentes codificações) e arquivos em HTML (diretamente da internet), além de produzir listas de frequência de palavras e concordâncias. O TextSTAT lê as páginas especificadas pelo pesquisador e as salva em um TextSTAT-corpus. Para fazer o download do TextSTAT basta acessar o link <http://neon.niederlandistik.fu-berlin.de/en/textstat/>.

Após este tratamento de frequência, nosso próximo passo será analisar os dados estatisticamente para se observar a associação entre pares de eventos. Para tal, faremos uso do teste do chi-quadrado. De acordo com Ellis e Ferreira-Júnior (2009), dentro da Lingüística de Corpus, um conjunto de medidas de associação como estas têm sido desenvolvidas para o caso particular de se determinar a co-ocorrência das palavras e outros elementos lingüísticos tais como construções. Dessa forma, este será o teste que usaremos em nossos dados.

A coleta, transcrição e análise do minicorpus embasará esta pesquisa, no sentido de que apresentará a capacidade de aprendizes não-nativos em usar tais construções/*chunks* formados pelo verbo *get* mais a partícula que indica o caminho/*path* explicitado em situações que expressam o cotidiano.

6 Considerações finais

Como nosso trabalho ainda está em fase de execução e análise, não temos dados concretos a serem apresentados, mas buscamos evidências que nos faça observar como se dá o uso de *chunks* com o verbo *get* por aprendizes/falantes de inglês como L2.

Por ser este um trabalho lingüístico, cuja preocupação de fundo é a aquisição de L2, todos os dados serão analisados e confrontados com teorias que tratem da hipótese central da interlíngua, ou seja, a transferência, assim como teorias que tratem da aquisição de construções/*chunks* – padrões recorrentes dos elementos linguísticos que servem a algumas funções linguísticas bem definidas. Também faremos uso de teorias relacionadas às línguas *satellite-framed* e *verb-framed*, teorias estas que embasam esta pesquisa, tudo isso seguindo a perspectiva da Linguística de Corpus, buscando, ao final, traçar algumas considerações e conclusões.

Referências bibliográficas

BEAUMONT, Digby; GRANGER, Colin. *English Grammar*: An Intermediate Reference and Practical Book. New Edition. Oxford: Ed. Heinemann, 1992. 352 p.

BIBER, Douglas *et al*. *Grammar of Spoken and Written English*. 7. ed. England: Ed. Longman, 2007. 1204 p.

BIBER, Douglas; CONRAD, Susan; LEECH, Geoffrey. *Student Grammar of Spoken and Written English*. 9 ed. England: Ed. Longman, 2010. 487 p.

CELCE-MURCIA, Marianne; LARSEN-FREEMAN, Diane. *The Grammar Book*: An ESL / EFL Teacher's Course. 2. ed. USA: Ed. Heinle & Heinle, 1999. 855 p.

CHANG, Fang; BAO, Yun-liang. Language chunks and college English writing. *Sino-US English Teaching*, USA, v.5, n.2, 2008. Disponível em: <http://www.linguist.org.cn /doc/su200802/su20080201.pdf>. Acesso em: 05 maio 2010.

ELLIS, N. C. Constructions, Chucking, and Connectionism: The emergence of second language structure. In: DOUGHTY, Catherine J.; LONG, Michael H. (Ed.). *The Handbook of Second Language Acquisition*. Oxford: Blackwell, 2003. p. 63-103.

ELLIS, N. C. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, v. 24, n. 2, p. 143-188, 2002a. Disponível em: <http://www.lotschool.nl/files/schools/archief/Winterschool%20Nijmegen%202007/dabrowska/Ellis%202002.pdf>. Acesso em: 05 maio 2010.

ELLIS, N. C.; FERREIRA-JUNIOR, F. Construction and their acquisition. Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, v. 7, p. 187-220, 2009. Disponível em: <http://web.mac.com/ncellis/Nick_Ellis/Publications. html>. Acesso em: 12 jun. 2010.

_____. Construction Learning as a Function of Frequency, Frequency Distribution, and Function. *Modern Language Journal*, v. 93, n. 3, p. 370-385, 2009. Disponível em <http://web.mac.com/ncellis/Nick_Ellis/Publications.html>. Acesso em: 12 jun. 2010.

LADO, Robert. *Linguistics across cultures:* Applied linguistics for language teachers. Ann Arbor: Michigan University Press, 1957.

MORAES BEZERRA, Isabel Cristina Rangel. Aquisição de Segunda Língua: de uma perspectiva lingüística a uma perspectiva social. *Revista SoLetras*, Departamento de Letras da Faculdade de Formação de Professores da UERJ, v. único, n. 5 e 6, 2003. Disponível em: <http://www.filologia.org.br/soletras/5e6/03.htm>. Acesso em: 28 Jul. 2009.

MURPHY, Raymond. *Essential Grammar in Use*: A self-study reference and practice book for elementary students of English. 2. ed. Cambridge: Cambridge University Press, 1997. 300 p.

MURPHY, Raymond. *Essential Grammar in Use*: A self-study reference and practice book for intermediate students of English. 3. ed. Cambridge: Cambridge University Press, 2004. 379 p.

NELSON, Peta L. *Grammar is Great*. 4. ed. Oxford: Ed. Heinemann, 1996. 256 p.

NUNAN, D. Case Study. In: _____. *Research Methods in Language Learning.* Cambridge: Cambridge University Press, 1992. p. 74-90.

_____. Introspective methods. In: _____. *Research Methods in Language Learning.* Cambridge: Cambridge University Press, 1992. p. 115-135.

ODLIN, Terence. Cross-Linguistic Influence. In: DOUGHTY, Catherine J.; LONG, Michael H. (Ed.). *Handbook of Second Language Acquisition*. Oxford: Blackwell, 2003. p. 436-486.

SAMPAIO, W.; SILVA,V.; SINHA, C. Espaço e Movimento em Amondawa: Violando a tipologia. *Pesquisa & Criação*, Porto Velho: PROPEX/EDUFRO, n. 4, p.130-136, Ago. 2005.

SARDINHA, Tony Berber. *Linguística de Corpus*. Barueri, SP: Manole, 2004. 412 p.

SELINKER, L. *Rediscovering Interlanguage*. 2. ed. New York: Longman, 1992. 288 p.

SNYDER, W. On the nature of syntactic variation: Evidence from complex predicates and complex word-formation. *Language*, v. 77, p. 324-342, 2001. Disponível em: <http://web2.uconn.edu/snyder>. Acesso em: 15 Set. 2010.

SNYDER, W.; LILLO-MARTIN, Diane. Motion Predicates and the Compounding Parameter. *Nanzan Linguistics*, Nagoya, Japan: Center for Linguistics, Nanzan University, n.2, p. 103-105, 2005. Disponível em : <http://web2.uconn.edu/snyder>. Acesso em: 15 Set. 2010.

TALMY, L. *Toward a cognitive semantics:* Concept structuring systems. Volume 1. Cambridge, MA: MIT Press, 2000.

_____. *Toward a cognitive semantics*: Typology and process in concept structuring. Volume 2. Cambridge, MA: MIT Press, 2000.

TAGNIN, Stella O. *O jeito que a gente diz*: Expressões convencionais e idiomáticas. São Paulo: Disal Editora, 2005. 120 p.

VINCE, Michael. *Language Practice*: Reference and Practice for Intermediate Students of American English. Oxford: Ed. Macmillan, 2000. 266 p.

_____. *Elementary Language Practice*. Oxford: Ed. Macmillan, 1999. 256 p.

Pacotes lexicais em corpus de aprendizes do ensino médio

Shirlene Bemfica de Oliveira[1]
Amanda Mendes de Oliveira Rossi[2]
Gabriela Maria Ferreira Leite[3]
Kamila Oliveira do Carmo[4]
Tatiane Morandi de Oliveira[5]

RESUMO: Este trabalho tem por objetivo mapear e descrever os pacotes lexicais (*lexical bundles)* típicos de alunos iniciantes evidenciados em um corpus de textos argumentativos. O estudo de caso foi desenvolvido com a participação da pesquisadora, quatro alunos bolsistas do Ensino Médio (bolsistas PIBIC Júnior) e aproximadamente 230 alunos da segunda série distribuídos em sete turmas do Ensino Médio de um Instituto Federal. Os dados foram coletados por meio da produção de um texto argumentativo escrito pelos alunos e as análises feitas com o auxílio das ferramentas *AntConc* e *Wordsmith Tools* e categorizados de acordo com a *Academic formulas List* (SIMPSON-VLACH; ELLIS, 2010). A análise quantitativa foi feita com base na freqüência dos itens investigados e de seus colocados para auxiliar na "compreensão do comportamento das palavras em determinados contextos de uso e freqüência, além de respaldar e enriquecer as análises" (BIBER, 1998, p. 8). Este tipo de análise

[1] Doutora em Linguística Aplicada, Professora de língua inglesa, IFMG – Campus Ouro Preto, shirleneo@yahoo.com

[2] Aluna do Ensino Médio Técnico em Edificações, IFMG – Campus Ouro Preto, (Bolsista CNPQ)

[3] Aluna do Ensino Médio Técnico em Edificações, IFMG – Campus Ouro Preto, (Bolsista CNPQ)

[4] Aluna do Ensino Médio Técnico em Edificações, IFMG – Campus Ouro Preto, (Bolsista IFMG)

[5] Aluna do Ensino Médio Técnico em Edificações, IFMG – Campus Ouro Preto, (Bolsista CNPQ)

possibilitou compreender melhor os padrões probabilísticos da linguagem produzida no contexto de uso e foi possível mapear as características do discurso típico de aprendizes iniciantes. A investigação das frequências dos traços linguísticos justifica-se, pois a comprovação da freqüência atestada é que levará o pesquisador a probabilidade teórica (BERBER SARDINHA, 2000; 2004).

PALAVRAS-CHAVE: pacotes lexicais, Linguística de Corpus, iniciantes, corpus de aprendizes

ABSTRACT: This work aims at mapping and describing the typical lexical bundles in beginner students' discourse in a learner corpus of argumentative texts. The case study was developed with the participation of the researcher, eight monitors (PIBIC Junior), and approximately 230 High School students from the second grade in a Federal Institute. The data was collected through essays written by the students and the analysis were done with the *AntConc* concordancer, *Wordsmith Tools* and categorized according to the *Academic formulas List* (SIMPSON-VLACH; ELLIS, 2010). The quantitative analysis was based on the investigated items frequency and their *collocates*, in order to help "in the comprehension of the behavior of the words in determined contexts of use and frequency, moreover to support and enrich the analysis" (BIBER, 1998, p. 8). This kind of research made possible to better understand better the language probabilistic patterns produced in the contexts of use and it was possible to map the features of the beginner students' typical discourse. The investigation of the frequency and linguistic traits is justified because the evidence of the attested frequency is the one which will take the researcher to the theoretical probability (BERBER SARDINHA, 2000; 2004).

KEYWORDS: lexical bundles, Corpus Linguistics, beginners, learner corpus

1 Introdução

O desenvolvimento da produção de textos nas aulas de língua inglesa é uma tarefa importante, porém muito árdua tanto para os alunos quanto para os professores. Os alunos iniciantes, geralmente, produzem textos com problemas de coesão provocados pela construção de períodos demasiado longos e com rupturas, repetições

lexicais além do uso escasso de conectivos e pausas mal elaboradas. Jácome e Gomes (2004) atribuem essa dificuldade a interferência da oralidade no discurso escrito. Segundo os autores, o professor deve reconhecer que, no processo de aquisição da habilidade escrita, o aluno parte da oralidade e transpõe para a escrita, o que torna a produção dele parcialmente incoerente. Segundo os autores, professor deve ensinar os alunos a utilizarem diferentes mecanismos de coesão porque a mera correção normativa e/ou ortográfica, não é subsídio suficiente para um aluno desenvolver sua escrita (JÁCOME; GOMES, 2004).

Pesquisadores interessados no processo de aquisição de línguas estrangeiras (LE) demonstram que esse fato é normal e faz parte do desenvolvimento linguístico do aprendiz. Estas pesquisas focam no mapeamento do processo de aquisição da LE, descrevendo e explicando o que o aluno produz enquanto está aprendendo a língua alvo, no caso deste trabalho o inglês. Estas amostras dão evidências do que os aprendizes sabem sobre a língua que estão aprendendo e do seu nível de desenvolvimento (ELLIS, 1997). Além da descrição e explicação das amostras produzidas pelos alunos, as pesquisas corroboram que é relevante analisar os fatores externos e internos implícitos no processo de aquisição. Um dos fatores externos é o contexto social em que a aprendizagem ocorre. (ELLIS, 1997, p. 05). As condições sociais influenciam nas oportunidades que os aprendizes têm de ouvir e falar a língua, bem como nas atitudes que eles tem em desenvolver sua interlíngua.[6] Outro fator externo é o insumo que os alunos recebem, ou seja, as amostras de língua estrangeira às quais são expostos; o aprendizado não pode ocorrer sem a ocorrência desses insumos (KRASHEN, 1983; ELLIS, 1997).

O processo de aquisição pode ser explicado em parte por esses fatores externos, mas precisamos considerar os fatores internos. Os aprendizes possuem mecanismos cognitivos que os capacitam a extrair informações sobre a língua alvo (inglês) do insumo para notar

[6] Interlíngua é um termo cunhado por Selinker em 1969 e se refere ao sistema de transição criado pelo aprendiz, ao longo de seu processo de assimilação de uma língua estrangeira. É a linguagem produzida por um falante não nativo a partir do início do aprendizado, caracterizada pela interferência da língua materna, até o aprendiz ter alcançado seu teto na língua estrangeira, ou seja, seu potencial máximo de aprendizado (SCHUTZ, 2006).

as regularidades e a sistematização desta língua. Eles também possuem a atitude de aprendizagem, o conhecimento prévio de como aprender uma língua e estratégias de comunicação que desenvolveram quando aprenderam sua língua materna (ELLIS, 1997).

É nesse sentido que o papel da intervenção pedagógica é crucial para aperfeiçoar esses processos cognitivos e promover insumos significativos para a aprendizagem. Os aprendizes seguem, segundo Ellis (1997, p. 13) um padrão de desenvolvimento particular por causa de suas faculdades mentais que são estruturadas de forma que "esta é a forma que eles têm que aprender". Essas faculdades regulam o que o aprendiz extrai do insumo e como ele estoca a informação em suas memórias. Nesse sentido, se a tarefa proposta estiver um pouco além do que o aprendiz está acostumado a fazer com eficácia (insumo + 1), ocorrerá algum tipo de processamento mental e o desenvolvimento linguístico. Isso demonstra a relação entre insumo compreendido, produção significativa (*output*) para a assimilação do *intake*, que é a porção de insumo que é seletivamente retirado do discurso para processamento futuro. A extração requer segmentação e seleção de partes da língua que são salientes (DOUGHTY, 2001, p.214). Portanto, a intervenção em sala de aula deve focar no desenvolvimento das habilidades de compreensão e produção oral e escrita (*reading, listening, writing, speaking*) em contextos e usos variados. Para que o aprendiz aprenda, ele tem que ser capaz de extrair informações gramaticais das sentenças, de modo a produzi-las em outros contextos e ou fazer inferências para mudar a organização das mesmas para expressar um sentido diferente (GREGG, 2001, p. 156).

No entanto, a descrição e explicação do processo de aquisição esbarram em questões metodológicas que devem ser consideradas. Por exemplo, que aspectos específicos da língua devem ser mapeados? O que significa dizer que um aluno aprendeu uma língua? A aquisição é definida em termos das manifestações de uso da língua pelo aprendiz que são comunicáveis com o nativo. Schmidt (1983) aponta padrões recursivos de aquisição em aprendizes do inglês como segunda língua. Em seu estudo de caso, os aprendizes (em contexto natural e de instrução formal) fizeram o uso recorrente de agrupamentos de palavras (*lexical bundles*), expressões fixas e fórmulas (*formulaic sentences*) para se comunicarem na língua alvo de forma produtiva. Por exemplo, um dos participantes utilizava formulas como "*Hi! How's it?, So, What's new? Can I have____?* para

se comunicar de forma eficaz, mas o estudo não consegue comprovar se o mesmo aprendiz conseguia utilizar o verbo *can* em outros contextos de uso. Outro problema apontado no estudo é tentar medir se no processo de aquisição foi considerado o sobre uso da forma linguística (se o aprendiz usa a mesma forma linguística e qualquer contexto mesmo que não seja apropriado) ou a omissão da forma linguística por insegurança ou outro fator. Essas fórmulas e agrupamentos de palavras são importantes e contribuem para a fluência do discurso não planejado. Uma implicação é o papel que essas fórmulas desempenham não somente melhorando o desempenho, mas também o processo de aquisição.

Outro ponto discutido por Schmidt (1983) é que além dos aprendizes desenvolverem a interlíngua pelo uso de formulas, expressões fixas e agrupamentos de palavras, eles também adquiriram a língua de forma sistemática. O estudo mostra a mesma sequência de desenvolvimento na aquisição e sugere que os aprendizes adquirem aspectos da língua seguindo rotas particulares de desenvolvimento, com as mesmas características adquiridas umas antes das outras.

Ellis (1997, p. 13) questiona o quão universal estes padrões de desenvolvimento são e se todos aprendem do mesmo jeito. Segundo o autor, a aquisição envolve diferentes tipos de aprendizagem. Por um lado, o aprendiz internaliza porções da estrutura da língua (*chunks*, *formulas*). Por outro lado, ele adquire as regras (conhecimento de uma dada característica linguística que é usada em um contexto e função particular). Em outras palavras Ellis (1997, p. 13) afirma que os aprendizes são engajados na aprendizagem do item e do sistema. Quando o aprendiz aprende a expressão fixa (*formula*) ele se engaja na aprendizagem do item. Eles aprendem a expressão como um todo não analisado. Quando ele aprende que partes da fórmula são acompanhadas por uma variedade de classes de palavras e que podem expressar uma gama de funções comunicativas, ele está se engajando na aprendizagem do sistema (regras em contextos). Então, a explicação do processo de aquisição envolve ambos os processos de aquisição do item e do sistema e como eles se inter-relacionam.

O trabalho que ora se apresenta foi motivado pela necessidade de compreender melhor o desenvolvimento da interlíngua de aprendizes de inglês como língua estrangeira. Ele tem como objetivo identificar e mapear os agrupamentos de palavras e seus colocados partindo do discurso de aprendizes iniciantes evidenciados em um

corpus escrito feito por eles. Neste estudo, temos a hipótese de que os grupos de alunos do Instituto Federal embora sejam heterogêneos em termos do perfil linguístico, eles estão em estágios do processo de aquisição da língua inglesa parecidos. Corroboramos os pressupostos teóricos acima relatados e acreditamos ser possível mapear as fórmulas e agrupamentos de palavras típicos do estágio de desenvolvimento deles. Faremos uma interface entre a Linguística Aplicada ao ensino de língua inglesa e a Linguística de Corpus para a compreensão dos dados compilados.

O estudo dos agrupamentos ou pacotes lexicais é importante, pois eles são sequências recorrentes em diversos registros (BIBER et al. 1999, p 13) e "são salientes devido à sua rigidez" o que os tornam bons padrões para o ensino e aprendizagem de uma Língua Estrangeira, pois são facilmente notados (SARDINHA, no prelo). Temos a hipótese de que os alunos do Instituto Federal embora estejam em grupos heterogêneos em termos do perfil lingüístico, eles estão em estágios do processo de aquisição da língua inglesa parecidos. Acreditamos ser possível mapear agrupamentos de palavras típicos do estágio de desenvolvimento deles, e para isso faremos uma interface entre a Lingüística de Corpus aplicada e o Ensino de Língua Inglesa com Foco na Forma para a compreensão dos dados compilados. Este trabalho também se justifica, pois as pesquisas sobre corpora de aprendizes são muito recentes e o caráter inovador deste tipo de estudo se deve a "uma grande carência de estudos sobre a interlíngua de aprendizes brasileiros" e de "compilações de corpus de aprendizes no Estado de Minas Gerais" (DUTRA, 2010, p. 03).

2 Referencial Teórico

Este artigo é embasado pela Linguística de Corpus (LC) que é a área do conhecimento que estuda a linguagem por meio da utilização do computador (GONZÁLES, 2007, p. 8). Ela é definida como uma maneira de se chegar à linguagem por meio da análise dos padrões probabilísticos que se constroem nos contextos em que os falantes os empregam (SARDINHA, 2000). A principal característica da LC é, segundo Gonzáles (2007, p. 8), a observação de dados empíricos armazenados em bancos de dados que compõem um corpus, com a utilização de ferramentas eletrônicas que auxiliam na análise de dados

verificando os fenômenos da língua em uso. Para este tipo de análise, recorremos a um corpus que é entendido como

> um conjunto de dados linguísticos (pertencentes ao uso escrito da língua), sistematizado segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso de algum de seu âmbitos, dispostos de tal modo que possam se processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SANCHEZ, 1995, p.8-9).

Entendemos que as unidades básicas da representação da linguagem (as palavras) são construções de mapeamentos de forma e significado, convencionados no discurso da comunidade e entrincheirado como conhecimento de língua na mente dos aprendizes (SIMPSOM-VLACH; ELLIS, 2010). Essas construções, segundo os autores, são associadas com funções semânticas, pragmáticas e discursivas particulares, e são adquiridas pelo engajamento na comunicação significativa. Essas construções formam um registro[7] estruturado do conhecimento do falante das convenções de sua língua, como unidades representadas na mente do falante. (SIMPSOM-VLACH; ELLIS, 2010, p. 1). A palavra,[8] como elemento básico da linguagem, é um construto muito complexo de se definir. Estudiosos de diversas áreas propõem formas de definir e categorizar o conceito. Na área de Lingüística de Corpus, a palavra é "relativamente fixa na sua forma interna, mas independente em sue papel em unidades maiores"; ela pode ser identificada pelo seu sentido ortográfico, gramatical ou léxico[9] (BIBER et al., 2002, 14).

As palavras podem ser agrupadas em três famílias de acordo com a função e comportamento gramatical: palavras lexicais,

[7] Registro: o estilo do texto em que a palavra é usada. No exemplo, "*President vows to support allies*", a ocorrência é mais provável em uma manchete de jornal, e "*vows*" no discurso oral se refere a "*marriages*"; o verbo "*vow*" é mais usado como "*promise*" (BIBER at al., 1999). Essa definição vai alem da noção de gênero textual.

[8] Todos os conceitos apresentado a seguir foram retirados e traduzidos da referência bibliográfica abaixo. BIBER, D. et al. Student Grammar of spoken and written English. London: Longman, 2002.

[9] Tradução minha para *lexeme*.

funcionais e *inserts.* As palavras lexicais são aquelas que carregam a informação do texto ou ato de fala. Elas podem ser divididas em classes ou partes do discurso (substantivos, verbos lexicais, adjetivos e advérbios). São chamadas de classes abertas e por isso são as mais numerosas os registros e frequentemente tem uma estrutura interna complexa que pode ser composta de várias partes (unfriendliness = un + friend + li + ness). As palavras funcionais (preposições, coordenadores, verbos auxiliares e pronomes) normalmente indicam a relação entre significados e nos ajudam a interpretar as unidades contendo as palavras lexicais mostrando com as unidades estão relacionadas. Elas pertencem a uma classe fechada, ocorrem frequentemente nos registros e tem um número limitado e fixo de agrupamentos. Os *inserts* são principalmente encontrados nos registros orais e não formam uma parte integral da estrutura sintática, mas tendem a ser inseridos naturalmente nos textos. Eles são marcados por quebras na entonação do discurso ou pela pontuação marcada na escrita. Eles carregam significados emocionais e discursivos e são usados para expressar a resposta emocional do falante a situação (*well, hum hm, cheers, bye, yeah, ugh*) (BIBER et al., 2002, 14-16).

As palavras também podem ocorrer em sequências combinadas como palavras múltiplas (*multi-words*), expressões idiomáticas e colocações. A palavra múltipla é uma seqüência de palavras ortográficas que funcionam como uma unidade gramatical única (preposição *on top of*, o advérbio *of course*). As expressões, como as palavras múltiplas, tem um significado que não pode ser previsto ou compreendido pelo significado dos seus constituintes separadamente, como em *fall in love* ou *make someone's mind*. E finalmente, a colocação que é a relação entre duas ou mais palavras lexicais independentes que aparecem juntas ou co-ocorrem frequentemente (BIBER et al., 2002, 18).

2.1 Pacotes lexicais

No caso deste estudo, o foco é dado nas palavras que co-ocorrem em sequências mais longas e frequentes, chamados de 'pacotes lexicais' (*lexical bundles*). De acordo com Biber et al. (1999), os pacotes lexicais são sequências de três ou mais palavras que mostram uma tendência estatística de co-ocorrerem juntas em determinados tipos de textos e, na maior parte dos casos, não são unidades estruturais completas (por exemplo, "a ver com" e "acordo

com a") nem expressões que os falantes reconheceriam como idiomáticas ou fixas. Os pacotes lexicais são definidos por sua frequência e pelos seus colocados,[10] e para serem considerados pacotes lexicais recorrentes, a combinação das palavras tem de ocorrer, pelo menos, dez vezes por milhão de palavras. Além disso, somente as combinações ininterruptas (não divididas por pontuação ou trocas de turno) podem ser tratadas como pacotes lexicais em potencial (BIBER et al., 1999). Os pacotes lexicais podem ser organizados de acordo com a função que desempenham no discurso como pacotes lexicais de opinião, como organizadores discursivos e e como pacotes lexicais referenciais (SIMPSOM-VLACH; ELLIS, 2010).

2.1.1 Pacotes lexicais de opinião

Os pacotes lexicais de opinião dão subsídios para a interpretação das proposições com dois sentidos: *epistêmico* e *atitudinal*. Os pacotes epistêmicos podem ser positivos e negativos e dão a ideia de certeza, incerteza, probabilidade e possibilidade. Eles comentam sobre o status do conhecimento da informação (BIBER et al., 2004). Os pacotes lexicais de opinião atitudinais expressam as atitudes dos falantes/escritores sobre as ações e eventos descritos nos exemplos. Eles podem ser pessoais e impessoais. Os pessoais são atribuídos aos falantes/escritores e os impessoais expressam exemplos similares sem serem atribuídos diretamente pelos falantes/escritores (BIBER et al., 2004).

Os pacotes lexicais são usados também para expressar as *atitudes* dos falantes/escritores sobre os eventos ou ações descritos nas proposições. Eles podem expressar desejo, obrigação, intenção e habilidade. Os pacotes que exprimem desejos incluem opiniões, mas também são usados para iniciar tópicos. Os pacotes que exprimem obrigação são diretivos e tem o pronome da segunda pessoa (*you*) como sujeito (BIBER et al., 2004, p 390). Os diretivos podem ser impessoais sem nenhum pronome pessoal, mesmo se a ordem for

[10] Os "colocados são as palavras que ocorrem ao redor do nódulo, em posições relativas. Difere de palavra de contexto porque esta é opcional, definida pelo usuário no momento da busca. Os colocados, contudo, são todas as palavras que ocorrem perto do nódulo, dentro do horizonte especificado, incluindo as palavras de busca que existirem" (SARDINHA, 2004, p. 188).

direcionada ao leitor/ouvinte. Eles são usados para prever intenções são em sua maioria pessoais e expressam as intenções do falante/escritor para uma ação futura. Estes pacotes são muito usados no momento das explicações (BIBER et al., 2004, p 391).

2.1.2 Organizadores Discursivos

Há pacotes lexicais que servem para introduzir, elaborar ou esclarecer tópicos. Quando desempenham a função de introduzir assuntos, eles geram o que Biber *et al* (1999) chamam de *syntatic blends.*[11] Segundo eles, o pacote sinaliza que o novo tópico está por vir, mas as duas partes não são bem formadas sintaticamente. Os pacotes que introduzem tópicos podem ocorrer na primeira ou na segunda pessoa e convidam, de acordo com o pronome usado, a desempenhar as ações das proposições.

A segunda subcategoria dos pacotes lexicais como organizadores do discurso auxiliam na elaboração e no esclarecimento do tópico. Segundo BIBER et al. (2004), os marcadores *you know* e *I mean* são usados como pacotes lexicais, normalmente quando o falante/escritor acredita que a explicação adicional é necessária. Os pacotes *as well as* e *on the other hand* são usados para explicitar comparação e contraste.

2.1.3 Pacotes lexicais referenciais

Os pacotes referenciais representam a maior categoria funcional e geralmente identificam entidades ou palavras. Ela descreve quatro maiores subcategorias incluindo: especificação de atributos, identificação e foco, contraste e comparação, dêiticos e locativos e marcadores de imprecisão (SIMPSON-VLACH; ELLIS, 2010, P. 17).

2.1.3.1 Pacotes de identificação e de foco

Estes pacotes são comuns no ensino na sala de aula, focalizando os substantivos das frases mais importantes no pacote. Por exemplo, o pacote *those of you who* identifica o subgrupo de estudantes que

[11] n. An utterance which switches part-way through from one well-formed structure to another, the whole being ill-formed; an example is *It's my car is the problem. (Dictionary of Grammatical Terms in Linguistics)

estão em foco. Exemplo: *For those of you who came late I have the, uh, the quiz.* E alguns casos, eles também tem a função de organizar o discurso. Esses pacotes são frequentemente usados depois de uma prolongada explicação para enfatizar ou resumir o principal ponto. Em outros casos, podem ser usados para introduzir o discurso declarando o primeiro ponto mais importante do texto e detalhando-o. (SIMPSON-VLACH; ELLIS, 2010).

2.1.3.2 Pacotes de imprecisão

A segunda maior subcategoria dos pacotes referenciais indica imprecisão. Estes duas funções específicas, ou indicar uma referência especificada que não é necessariamente exata, ou indicar que há referências adicionais do mesmo tipo que poderiam ser usadas.

2.1.3.3 Pacotes de atribuição específica

Identifica os atributos que segue *head noun* (núcleo do sintagma nominal). Alguns desses pacotes especificam quantidade. Exemplo: You´d *have a lot* of power (classroom teaching). O pacote *little bit* usualmente especializa na função de introduzir o tópico, aparentemente minimiza a expectativa requerida dos estudantes. Exemplo: So I want to talk *a little bit* about process control from that point of view (SIMPSON-VLACH; ELLIS, 2010). Outros pacotes têm a função descrever o tamanho e a forma do head noun. Exemplo: *These figures give an idea of the size of the ethnological community in Russia*. (textbook). Os pacotes abstratos são usados também para estabilizar a relação lógica no texto. Eles são definidos em termos das emoções que suscitam.

2.1.3.4 Pacotes que indicam dêiticos, tempo e lugar

Muitos pacotes referenciais remetem a lugares particulares, tempos ou locações no próprio texto. Exemplo: *Children in the United States are not formally employed in farm work* (...) (textbook). Os dêiticos são comuns no registro escrito, onde fazem referencia direta a figuras do texto. Exemplo: *As shown in figure* 4.4, the higher the real (...) Muitos destes pacotes são multifuncionais, referindo a lugar, tempo, texto deixis, dependendo do contexto. Exemplo: So you have to

Record that, since the asset was sold at the end of the year (classroom teaching)

2.2 Variação do registro na exploração funcional dos pacotes lexicais

Há uma variação no uso e nas funções desempenhadas pelos pacotes lexicais como um todo. Os pacotes de opinião são extremamente comuns na sala de aula e na conversação. Os organizadores discursivos são comuns na sala de aula e menos comuns na conversação. Os pacotes referenciais são extremamente comuns na sala de aula e menos comuns em textos e conversas acadêmicas (SIMPSON-VLACH; ELLIS, 2010). Os padrões usados na sala de aula ajudam na expansão porque o pacote lexical é geralmente muito comum nesse registro. Na sala de aula, combinam características da conversação (usando opiniões e pacotes organizadores do discurso). Contudo, na sala de aula atualmente usam todos esses três tipos em maior medida do que em outros registros (SIMPSON-VLACH; ELLIS, 2010).

Existem dois padrões comuns em sala de aula: combinam a função e a prioridade comunicativa de envolver no discurso falado (demonstrado pelo uso de feixes densos referenciais). Além disso, a sala de aula é uma estrutura com pacotes lexicais em maior medida do que em qualquer outro registro. Isso mostra um grande número de pacotes lexicais organizadores do discurso usados na sala de aula (muitos referenciais são usados para uma função do discurso, tal de identificação/foco, imprecisão, e especificação quantitativa). O padrão aparente mostra a complexidade de comunicar nesse registro e os pacotes lexicais são úteis para instrutores porque tem a necessidade de organizar e estruturar o discurso uma vez informar, envolver e produzir em tempo-real restrições de produção (SIMPSON-VLACH; ELLIS, 2010).

2.3 A relação entre as categorias funcionais e estruturais

Há uma relação forte entre o tipo estrutural e função do discurso dos pacotes lexicais. A maioria dos pacotes de opinião são compostos de sintagma nominais (*noun phrases*) ou fragmentos de frases preposicionais (*prepositional phrases*). A predicação/intenção dos

pacotes de opinião é composta de fragmentos de sintagmas verbais (*VP-based bundle*), principalmente a incorporação do semi-modelo *be going to*. Os organizadores discursivos têm uma categoria funcional que usa todos três tipos de estrutura. Esse padrão é fortemente associado com o registro: conversação, 'oral', usa principalmente sintagmas verbais (*VP-based*) e pacotes lexicais de orações dependentes (*dependent clause lexical bundles*) para a função de opinião.

A conversa acadêmica, 'literatura', usa pacotes lexicais baseados em sintagmas nominais e sintagmas preposicionados para as funções referenciais. Na sala de aula é parecido geralmente com as características gramaticais, e usa o mesmo tipo de pacote lexical. Os padrões sugerem uma associação direta entre forma e função para os pacotes lexicais. Exemplo: pacotes baseados em orações complementares são usados com a função de dar opinião. Assim, usa sentido de que é muito comum multi-palavras sequenciais com complemento de orações, se tornaria fixa a estrutura dos pacotes lexicais de opinião. Similarmente, sintagmas nominais e preposicionados (*noun phrase* e *prepositional phrases*) são dispositivos da primeira gramática que usam funções referenciais, e fazem sentido comumente nas multi-palavras sequenciais se tornaria fixa em pacotes referenciais. Assim, é complexa a interação entre forma estrutural, função do discurso, efeitos típicos e características situacionais do registro (SIMPSON-VLACH; ELLIS, 2010).

2.4 O status teórico dos pacotes lexicais

Para Simpson-Valch e Ellis (2010), a análise e categorização dos pacotes lexicais desta forma sugerem que eles devem ser considerados como uma construção linguística básica com funções importantes na construção de um discurso. Entretanto, com relação a estrutura e função, eles diferem dramaticamente de outras características linguísticas. Levando em consideração que eles são definidos estritamente com base na frequência, sem consideração de critérios estruturais ou funcionais, eles podem ser sequências arbitrárias de palavras que não tem status linguísticos. Em vez disso, essas palavras frequentes acabam sendo facilmente interpretadas tanto em termos estruturais quanto em termos funcionais. Essas sequências de palavras podem ser consideradas como quadro estrutural, seguido por uma abertura. O quadro de funções mostra

como uma espécie de discurso âncora para a nova informação contando ao ouvinte/leitor como interpretar essa informação com respeito ao status de postura, organização de discurso ou referencial. O uso dos pacotes lexicais são notadamente diferentes daqueles encontrados em características das gramáticas tradicionais.

Podemos perceber com essa categorização que o ensino em sala de aula que mistura características orais e literais no uso dos pacotes lexicais, na verdade, vai além dos alvos esperados em seus padrões de uso. Ele mostra um uso mais extensivo de pacotes lexicais e pacotes organizadores discursivos. E apontam para a importância que os pacotes lexicais têm de suas frequências de uso e funções do discurso óbvio (SIMPSON-VLACH; ELLIS, 2010). Certamente, outras abordagens com objetivos diferentes são importantes para complementar esse estudo e aumentar a compreensão de "*multi-words*", incluindo estudos de psicolinguística e investigações de mais expressões idiomáticas perceptualmente salientes e ao mesmo tempo, este estudo ilustra como uma abordagem exploratória de corpus facilita a identificação de características de linguagem que passam despercebidos de outra forma, mas que acabam por ser uma parte fundamental para escritores e oradores em repertório comunicativo (SIMPSON-VLACH; ELLIS, 2010).

3 Metodologia

O estudo, de natureza empírica, refere-se a um estudo de caso desenvolvido com a participação de aproximadamente 230 alunos da segunda série distribuídos em turmas do Ensino Médio de um Instituto Federal. As turmas, de segundo ano do nível básico (ano base 2010), tinham um encontro semanal (1h e 40 min.) e utilizam o material didático *Straight Foward Elementary*. Durante as aulas eram desenvolvidas atividades que contemplavam as habilidades de compreensão e produção oral e escrita (*listening, speaking, Reading and writing*), além de pronúncia, gramática e vocabulário.

Os dados foram coletados em três fases. Na primeira fase, foi solicitado aos alunos que escrevessem um texto argumentativo sobre distúrbios alimentares (150-180 palavras). Os alunos trabalharam em grupos de 3 ou 4 componentes e receberam a instrução de que o texto deveria ser constituído por parágrafos curtos, início e desenvolvimento do assunto, opiniões, argumentos e exemplos. Eles

foram orientados a escrever argumentos baseados em uma vida saudável e uma conclusão que respondesse ao primeiro parágrafo ou simplesmente com a ideia chave da opinião deles. Nesta fase, o corpus foi digitado pelos alunos bolsistas e nomeado *corpus 1*.

Na segunda fase, a pesquisadora e os bolsistas utilizaram dois concordanciadores para gerar as listas de palavras mais freqüentes, para a identificação dos pacotes lexicais e para gerar as linhas de concordância: o *AntConc* e o *Wordsmith Tools*. Os dois foram usados, pois o grupo somente teve acesso as versões gratuitas disponíveis *online*. Após a indexação da primeira produção escrita, a pesquisadora (professora das turmas) fez uma intervenção pedagógica com foco na forma por meio de atividades de leitura, compreensão e produção oral com foco nas orações relativas. Na fase final da pesquisa, os alunos reescreveram os textos participando dos mesmos grupos enriquecendo o conteúdo e corrigindo os possíveis problemas. Para a reescrita, os textos foram lidos pela professora e erros foram marcados (ortografia, concordância, ordem de palavras, omissão de palavras, etc). O *corpus 2* foi digitado e compilado pelos bolsistas. As análises foram feitas com cunho quantitativo e qualitativo. A análise quantitativa foi feita com base na freqüência dos itens investigados e na categorização proposta por Simpson-Vlach e Ellis, 2010. O quadro 01 abaixo apresenta as tarefas e os objetivos respectivos:

QUADRO 1
Tarefas e respectivos objetivos

FASES	TAREFAS	OBJETIVO
Fase 1 Corpus 1	Produção de texto 1 Distúrbio alimentar	- Verificar a ocorrência de pacotes lexicais (*lexical bundles*) nas produções de textos dos alunos iniciantes.
Fase 2	Atividades com foco na forma	- Aumentar a incidência de *noticing* - Dar evidência positiva
Fase 3 Corpus 2	Produção de texto 2 Reescrita do texto 1	- Verificar as mudanças nas produções dos alunos iniciantes.
Análise	Análise comparativa das produções de texto	- Contrastar o discurso produzido pelos alunos antes e depois da atividade com foco na forma

Os pacotes lexicais foram analisados em linhas de concordância que apresentavam os itens específicos, dispostos de modo que seu contexto original foi mantido (SARDINHA, 2004, p. 187). Foi feita a análise da prosódia semântica dos pacotes considerando a conotação que as palavras carregam.

4 Análise de Dados

A investigação para este trabalho baseou-se em dois corpora sobre distúrbios alimentares. O primeiro com 41 textos e o segundo com 40 textos. Para a observação da macroestrutura do corpus, fizemos a análise da composição do corpus, dos títulos apresentados nos textos e geramos uma lista das palavras mais frequentes. A maioria dos textos tinha um padrão: definição do problema, apresentação das causas e consequências, exemplos e argumento baseados em uma vida saudável. Essa padronização é decorrente da instrução que receberam para a produção do texto.

O quadro abaixo mostra o tamanho dos corpora comparados em relação ao número total de palavras (12290 *word tokens*) e o número de palavras usadas nos textos (1925 *word types*) em 804 sentenças analisadas. A riqueza lexical, segundo Finatto et. al. (2011), "corresponde a uma medida dada pela razão entre o número de palavras diferentes (formas) existentes no corpus com o número total de palavras" (p. 219).

QUADRO 2
Composição do corpus (Wordsmith Tools)

Text File	File Size	Tokens	Types	Type/Token	Sentences
Overall	74311	12290	1925	15,7	804
corpus1.txt	35536	5903	1375	23,3	395
corpus2.txt	38775	6387	1456	22,9	409

Percebemos que pela relação *token* / *type*, os textos dos alunos não tiveram uma riqueza lexical e atribuímos isso ao nível lingüístico deles e ao vocabulário reduzido que dispunham para a produção escrita. Alem disso, podemos observar, pelo quadro, um aumento reduzido do número de palavras do primeiro corpus para o segundo. Acreditamos que o aumento somente aconteceu devido ao tipo de instrução que

eles receberam em relação ao erro. A tendência na maioria dos textos foi simplesmente corrigir os problemas ortográficos e gramaticais como mostra a linha de concordância abaixo.

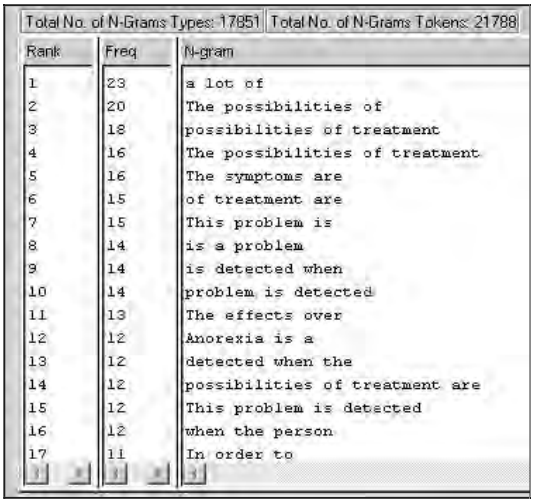
Corpus 1: *This problem is detected by isn of weight excessive the*
Corpus 2: *This problem is detected by losing weight excessively in a little*

Em relação aos títulos, no corpus 1 a maioria dos textos receberam o nome do distúrbio como título, por exemplo: *bulimia, anorexia, night time eating syndrome, childhood obesity, obesity, diabetes, anemia*. Quatro textos receberam outros títulos, como: *Finding a perfect body, Obesity: the big problem, Bulimia: the terrible disease, Bulimia: a food disturbance*. Um texto recebeu o título de *Introduction* e 5 textos ficaram sem títulos. No corpus 2, quatro textos continuaram sem títulos e a maioria manteve o nome do distúrbio como título.

A respeito da frequência das palavras, observamos que as palavras gramaticais foram as mais freqüentes como em todos os textos deste registro. As primeiras palavras gramaticais que apareceram foram: *the* (749), *and* (493), *to* (354), *a* (350) e *is* (334). *the* a segunda mais frequente é um determinante (*determiner*) que marca os substantivos como referentes de algo ou alguém que assume ser os falantes, leitores ou escritores (BIBER, 2002, p. 70). Após as gramaticais, a primeira palavra lexical é *problem(s)* (212) seguida das palavras *people* (126), *obesity* (91), *eating* (89), *eat* (79), *treatment* (79), *anorexia* e *weight* (72), disease (62), health (44), body e fat (43), family e symptoms (40), que foram bem relacionadas com os assuntos dos textos.

4.1 Análise dos pacotes de três e quatro palavras

Buscamos com o *AntConc* os pacotes lexicais de três palavras quatro palavras e obtivemos o número de 17851 *N-Grams Types* e 21788 *N-Grams Tokens*. O quadro abaixo mostra os dezessete mais freqüentes:



Total No. of N-Grams Types: 17851 | Total No. of N-Grams Tokens: 21788

Rank	Freq	N-gram
1	23	a lot of
2	20	The possibilities of
3	18	possibilities of treatment
4	16	The possibilities of treatment
5	16	The symptoms are
6	15	of treatment are
7	15	This problem is
8	14	is a problem
9	14	is detected when
10	14	problem is detected
11	13	The effects over
12	12	Anorexia is a
13	12	detected when the
14	12	possibilities of treatment are
15	12	This problem is detected
16	12	when the person
17	11	In order to

Figura 1: Pacotes lexicais de três e quatro palavras

Observamos que os pacotes mais frequentes eram compostos de sintagmas nominais, preposicionados ou de pacotes lexicais desempenhando diferentes categorias funcionais. Por isso, decidimos agrupá-los de acordo com a classificação proposta por BIBER et *al.* (2004); Simpson-Vlach e Ellis (2010) como expressões de opinião, organizadores discursivos e expressões referenciais. Em seguida, discutimos alguns exemplos separadamente.

4.2 Pacotes lexicais de opinião

Nos dados Neste estudo, utilizamos a categorização O quadro abaixo mostra as ocorrências dos pacotes que exprimem opinião.

QUADRO 3
Expressões de opinião

<table>
<tr><th>A. Opinião epistêmica</th><th colspan="2">B. Opinião Atitudinal</th></tr>
<tr><td>a1. pessoal</td><td>b1 desejo</td><td>b3. intenção / previsão</td></tr>
<tr><td rowspan="2">to think that (2)
what she thought (2)
people (that) think (4)</td><td>if you want to avoid (1)
people want a (1)</td><td>pessoal</td></tr>
<tr><td>b2 obrigação / diretivo</td><td rowspan="2">impessoal</td></tr>
<tr><td>a.2 impessoal</td><td>pessoal</td></tr>
<tr><td rowspan="2">according to (1)</td><td rowspan="3">people do not have to eat (1)
look at mirror (1)</td><td>b4 habilidade</td></tr>
<tr><td rowspan="2">pessoal</td></tr>
<tr><td>Hedges</td></tr>
<tr><td>may be (7)</td><td>impessoal</td><td>impessoal</td></tr>
<tr><td>kind of (2)</td><td>it is necessary to (3)</td><td>can be used to help (2)</td></tr>
</table>

4.2.1 Opiniao Epistêmica

A maioria dos pacotes lexicais que exprimem opinião no banco de dados é construída pelos verbos *think* e *know*. Eles são pessoais e, na maioria das vezes, dão a idéia de incerteza. No exemplo do corpus, *the parents do not know that these games are bad for the children* o pacote '*do not know that'* expressa somente incerteza. Agora, os pacotes com *(don't) think* apresentados no quadro abaixo, expressam possibilidade, mas uma falta de certeza, *when the people think are fat and*, *depression to think that was*, *with people that think they are fat.*

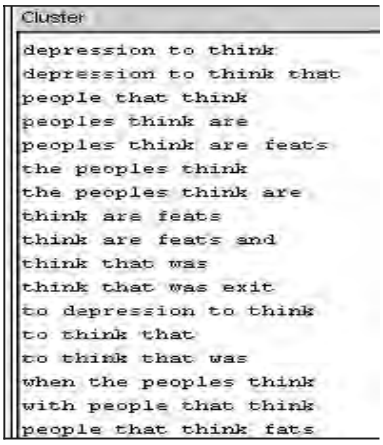


Figura 2: pacote lexical com o verbo *think*

É importante observar que esses pacotes assim categorizados podem também desempenhar outras funções. Eles podem servir para identificação referencial ou para introduzir novos tópicos. Todavia, os pacotes epistêmicos impessoais em contraste com os pessoais, expressam mais graus de certeza do que incerteza (BIBER et al., 2004, p 389). O quadro abaixo mostra ocorrências do pacote lexical que exprime opinião atitudinal impessoal *it is necessary to.*

4.2.2 Opinião Atitudinal

Os pacotes lexicais atitudinais foram usados para dar sugestões para solucionar os distúrbios alimentares, para iniciar tópicos e expressar opiniões, como mostra a figura abaixo:

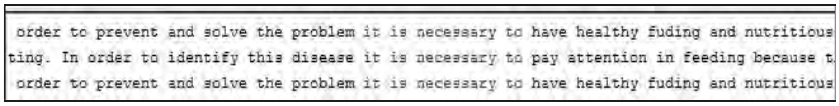

order to prevent and solve the problem it is necessary to have healthy fuding and nutritious
ting. In order to identify this disease it is necessary to pay attention in feeding because t
order to prevent and solve the problem it is necessary to have healthy fuding and nutritious

Figura 3: pacote lexical *it is necessary to*

Em nosso corpus, consideramos também os pacotes em que o verbo *have to* foi utilizado com a terceira pessoa para sugerir mudanças. Os diretivos, no geral muito usados na explicação, não foram muito recorrentes no corpus devido a pouca fluência dos alunos na língua e daí a dificuldade em explicar e argumentar sobre os problemas propostos.

4.3 Organizadores Discursivos

No corpus, os pacotes lexicais usados para introduzir os assuntos foram usados para dar exemplos, sugestões, para apresentar causas, conseqüências e para argumentar sobre os problemas discutidos. O quadro abaixo apresenta algumas ocorrências desses pacotes lexicais:

QUADRO 4
Organizadores Discursivos

Metadiscurso e referencia textual	**b. elaboração do tópico / esclarecimento**
the article is about (2) the text is (2)	in order to (13) And if you (6)
Introduçãodo tópico / foco	in other words (2)
the possibilities of treatment (17) this problem is detected (13) according to the (4) the best alternative is (2) although there is no (1)	

O pacote *in order to* ocorreu 13 vezes no corpus, na maioria dos casos com a função de sugerir mudanças ou solucionar o problema descrito. Neste caso, a configuração foi usada para iniciar a sugestão com a preposição in + substantivo order + preposição to seguido de um verbo. As três palavras tem o sentido de para (*for the purpose of*) e não são compreendidas separadamente. Houve uma ocorrência em que após o pacote lexical, os alunos colocaram um substantivo no lugar do verbo *in order to the organism not feel the necessity of any vitamin, or protein.* No exemplo, *in order to* tem o sentido de 'para que o organismo não sinta". Esta construção não foi encontrada em nenhum corpus de referencia da língua inglesa. Atribuímos o uso a uma transferência da língua materna. O quadro abaixo mostra algumas das ocorrências do *cluster*:

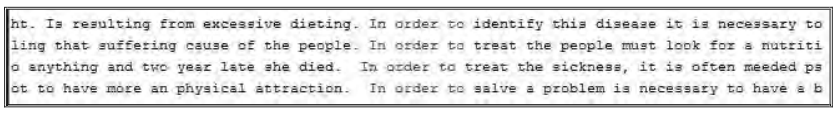

ht. Is resulting from excessive dieting. In order to identify this disease it is necessary to
ling that suffering cause of the people. In order to treat the people must look for a nutriti
o anything and two year late she died. In order to treat the sickness, it is often meeded ps
ot to have more an physical attraction. In order to salve a problem is necessary to have a b

Figura 4: pacote lexical *in order to*

A segunda subcategoria dos pacotes lexicais organizadores discursivos auxiliam na elaboração e no esclarecimento do tópico. Segundo BIBER et al. (2004), os marcadores *you know* e *I mean* são usados como pacotes lexicais, normalmente quando o falante/ escritor acredita que a explicação adicional é necessária. Os pacotes *as well as* e *on the other hand* são usados para explicitar comparação e contraste.

4.4. Pacotes Referenciais

Os pacotes referenciais geralmente identificam uma entidade ou seu atributo como importantes. BIBER et al. (2004) categorizam quatro subcategorias desse tipo de construção: identificação / foco, indicador de imprecisão, especificação de atributo e tempo/lugar/ referência textual.

QUADRO 5
Pacotes lexicais referenciais

<table>
<tr><td colspan="3">1. ESPECIFICAÇÃO DE ATRIBUTO</td></tr>
<tr><td>a) Atributo Intangivel</td><td colspan="2">b) Atributo Tangível
forma / tamanho/ quantidade</td></tr>
<tr><td>the problem is (12)
based on (5)
this case (5)
kind of (4)
according to (4)
focus on (2)
(in) the case (of) (3)
in which (3)
the development (of) (1)
to the extent that (1)</td><td>a lot of (23)
increase of (8)
the most serious (5)
high blood pressure (5)
and change of (4)
there are people (4)
and many others (3)
in most cases (3)
corporal mass rate (3)
a little time (2)</td><td>a little fat (2)
a little city (2)
daily rate (2)
the value of (2)
there’s no (2)
some / any types of (2)
has lot of (1)
with a lot of (1)</td></tr>
<tr><td rowspan="4">2. IDENTIFICAÇÃO / FOCO
is a (72)
such as (12)

this problem is (23)
with people who (12)
(she/they) do/did not (7)
an/one example of (5)
it is not (4)
people who commit (4)
he/she/it was a (3)

it can make (2)
before that this (2)
that we know (2)
which is (2)
have to do (1)
referência textual
in may 2008 (6)
on TV (12)</td><td colspan="2">3. CONTRASTE E COMPARAÇÃO
Related to (5)
Associated with the (4)
Different from (2)</td></tr>
<tr><td colspan="2">4. DEITICOS E LOCATIVOS
in the (47)
in the world (5)
at the (4)
in May 2008 (2)
in a little time (2)
in the U.S.A (1)</td></tr>
<tr><td colspan="2">5. INDICADORES DE IMPRECISÃO
problem like this (12)
some problem like that (1)
two or more people (1)</td></tr>
<tr><td colspan="2">abstrato
the result can be (2)
in the case of being (1)</td></tr>
</table>

O pacote lexical de identificação e foco mais freqüente foi *"is a"*. As ocorrências aparecem no início dos textos quando os alunos definem e explicam o problema e ocorrem em duas variedades de padrão: padrão de caracterização e padrão de identificação (BIBER et *al.*, 1999).

Upward collocates: 'is a'
variação: a, an
substantivos: Bulimina is a disease
adjetivos seguidos de substantivos: Bulimia is a psychological disturbance
downward collocates: sempre substantivos

Ambos os padrões contem o verbo cópula (is) e um predicativo expressando o papel semântico de atributo (BIBER et *al.*, 1999, 145). Segundo os autores, as sentenças com padrão de caracterização, tem um participante caracterizado como sujeito e respondem as perguntas 'O que é?', 'Como é?', 'Como mudou?' Nos exemplos dos alunos, a maioria das sentenças segue este padrão no momento em que eles definem o problema, como mostram as linhas de concordância abaixo:

bulimia is a problem related to food compulsion
anorexia is a common problem among population where people make
bulimia is a disease in which the person eat big quantity of food
obesity is a serious problem, caused by the excess of food
bullying is a global problem which can occur anywhere in which
it is a very serious problem and affect

Neste tipo de oração, a propriedade é atribuída ao referente do sujeito. A propriedade pode ser expressa por um sintagma nominal ou adjetivo. O padrão de identificação responde a pergunta 'Qual é?', 'Quem é?'e é formado pelo verbo cópula 'is' e um predicativo do sujeito. O predicativo é um sintagma nominal definido e não um sintagma adjetivo ou sintagma nominal indefinido, como é normalmente o caso do padrão de caracterização (BIBER et *al.*, 1999, 146). Este padrão expressa a identidade entre o sujeito e o predicativo. No banco de textos, o padrão foi recorrente quando os alunos narravam ou apresentavam exemplos dos problemas, como mostram as linhas de concordância.

one example for this case is an adolescent who is a nineteen years old girl
José, 17years old, play basketball. He is a student and suffers from obesity
one example of the problem is a girl whose name is Miriam
Manoela is a 16 years old teen, who had the dream to be a top

A análise feita acima é importante, pois nos leva a compreender melhor a estrutura da explicação, que segundo Sinclair (1991) leva as hipóteses sobre inferências, metalinguagem e a natureza geral da afirmativa lexical. Na primeira parte da explicação temos o tópico da sentença e seu contexto. A segunda parte é um comentário explicatório ou definitório seguido dos operadores.

No corpus, o pacote lexical mais freqüente com 23 ocorrências é o *a lot of* que aponta os atributos da palavra chave especificando quantidades. Ele foi usado para introduzir tópicos e para estabelecer relações lógicas no texto. No entanto, ao analisarmos todas as ocorrências mostradas na barra de rolagem, percebemos que ele também é associado lexicalmente em diferentes pacotes. a lot of (mais 71 ocorrências), a lot (77 ocorrências) e lots of (40 ocorrências).

about the subject and after, analyze a lot of cases choosing one
or this case is: Drink a lot of liquids. Replace sugar for sweetener
Eat a lot of fiber and fruits. Practice exercises.
A lot of soap opera show people
we can say a lot of examples today
got a lot of complications, died in the
has a lot of different causes, among than are culture, anxiety
ate a lot, so after she vomits all that
but without lots of repercussion,
a diabetic and has lots of problems with circulation

O pacote *a lot of* é categorizado como determinante que quantifica o pronome ou os substantivos contáveis e não contáveis. BIBER et al (2004, p. 278). Afirma que *a lot of* é mais comum na conversação e quase não ocorre em textos acadêmicos acreditamos que em nossos textos a ocorrência é alta devido ao gênero artigo de revista que também tem uma ocorrência alta nos dados de Biber.

Os *dêiticos* têm por objetivo localizar o fato no tempo e espaço sem defini-lo. Alguns pronomes demonstrativos podem ser expressões dêiticas bem como certos advérbios. No corpus, o pacote lexical *in the*, foi utilizado 47 vezes, precedido de sintagmas nominais (adjetivos + substantivos) ou substantivos.

a problem that affect many children in the world. Obesity is
very common in the adolescence, mostly in girls that desire a
body can exaggerate a lot of times. In the world we live today
is beautiful, in the fashion world where the models are skinny.
she suffered during a period of life. In the beginning of the suffering
a paranoid. They look themselves in the mirror, and their mind
Nowadays he is well and live happy in the USA. Some people are affected by
The effects over health are problems in the digestive system , points often
Height and Divide weight by height, in the case of being above the standard

5 Conclusão

Este artigo demonstra a possibilidade de desenvolver pesquisas envolvendo alunos do Ensino Médio na área de Linguística de Corpus, como participantes e como pesquisadores. Os bolsistas compreenderam aspectos relativos à área, foram capazes de organizar o corpus, usar os concordanciadores e de identificar padrões de acordo com a categorização proposta por Simpson-Vlach e Ellis (2010). Além disso, receberam insumos significativos da língua em uso e tiveram a oportunidade de interagir na língua alvo o que pode ter contribuído para o desenvolvimento da interlíngua deles e para a conscientização linguística.

Em relação à metodologia utilizada, percebemos a relevância de dar oportunidade, mesmo para os alunos iniciantes, de usarem a LE para produzir textos. Apesar de os textos apresentarem uma linguagem truncada, pudemos perceber o conteúdo apropriado ao tema (distúrbios alimentares), as causas e consequências e alguns mecanismos linguísticos para orientar o leitor para as possibilidades de cura ou prevenção dos distúrbios. No entanto, a metodologia foi limitada, pois não foi possível verificar o desenvolvimento linguístico dos aprendizes pelo fato de a coleta ter acontecido por um período muito curto. Todavia, percebemos a ocorrência de pacotes lexicais nos textos dos alunos. Apesar de terem um vocabulário reduzido e pouco conhecimento gramatical, houve a produção de pacotes lexicais desempenhando as mesmas funções categorizadas por Simpson-Vlach e Ellis (2010).

Os resultados aqui apresentados precisam ser ampliados através da exploração da frequência de pacotes lexicais em produções de textos de alunos mais avançados do Ensino Médio, através da expansão do corpus desta natureza, bem como a inclusão de outros registros para comparação. Alem disso, seria louvável a comparação dos resultados com corpora de referência.

Referências Bibliográficas

BERBER SARDINHA, A. P. Linguística de Corpus: Histórico e Problemática. In: D.E.L.T.A. v.16, n. 2, 2000, p. 323-367.

BERBER SARDINHA, A. P. Pesquisa em Lingüística de Corpus com WordSmith Tools no prelo.

BERBER SARDINHA, A. P.Linguística de Corpus. Barueri-SP. Manole, 2004.

BIBER, D. et al. Grammar of spoken and written English. London: Longman, 1999.

BIBER, D. et al. If you look at…: Lexical bundles in University teaching and textbooks. In: Applied Linguistics. Oxford: Oxford University Press. 25/3: 371-405, 2004

BIBER, D. et al. Student Grammar of spoken and written English. London: Longman, 2002.

BIBER, D. et al. Grammar of spoken and written English. London: Longman, 1999.

BIBER, D. S.; CONRAD; REPPEN, R. Corpus Linguistics: Investigating language structure and use. Cambridge: Cambridge University Press, 1998.

BOGDAN, R. BIKLEN, S. Investigação Qualitativa em Educação. Uma introdução à teoria e aos métodos. Tradutores: ALVAREZ, M. J. SANTOS, S. B. BAPTISTA, T. M. Portugal: Porto Editora, 1994.

BROWN, J. D. RODGERS, T. Doing second language research. Oxford: Oxford University Press, 2002, p.21-78.

Dictionary of Grammatical Terms in Linguistics. Disponível em: http://www.bookrags.com/tandf/syntactic-blend-tf/ Acesso em 30/06/2011.

DUTRA, D. P. Agrupamentos lexicais na escrita de aprendizes brasileiros de inglês: um estudo baseado em corpus. Plano de trabalho apresentado ao Programa Pesquisador Mineiro. Edital FAPEMIG 03/2010.

DUTRA, D. P. Conscientização linguística com base em corpora online. Intercâmbio XX, p. 79-98, 2009.

DUTRA, D. P.; SILERO, R. P. O uso de for: uma análise de itens linguísticos em corpus de aprendizes. Trabalho apresentado no VIII Encontro de Linguística de Corpus- RJ – UERJ. 2009

GONZÁLES, Z. M. G. Lingüística de Corpus na análise do Internetês. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) - Faculdade de Letras. Pontifícia Universidade Católica de São Paulo, São Paulo, 2007.

JÁCOME, A. J. P. C. A.; GOMES, N. S. A produção do texto narrativo na escola: influências da oralidade ou modalidade sintática? In: Revista Philologus. Ano 10, n. 28, 2004. Disponível em: http://www.filologia.org.br/revista/artigo/10(28)03.htm Acesso em 15/04/2010.

LANGACKER, R. W. Foudations on Cognitive Grammar. Descriptive applications. Standford, CA: Standford University Press, 1987.

SCHMIDT, R. Interaction, acculturation and the acquisition of communicative competence: a case study of na adult. In: WOLFSON, N.; JUDD, E. (eds.). Sociolinguistics and Second Language Acquisition. Newbury House, 1983, p. 168-169.

SCHÜTZ, R. Interferência, interlíngua e fossilização. Publicado em: 2006 Disponível em: http://www.sk.com.br/sk-interfoss.html Acesso em 26/04/2010.

SIMPSON-VLACH, R.; ELLIS, N. An academic formulas list: new methods in phraseology research. In: Applied Linguistics. Advance Access Published January 12, 2010, p. 1-26.

SINCLAIR, J. Corpus, concordance and collocation. Oxford: Oxford University Press, 1991.

Analisando um corpus oral de aprendizes: um estudo comparativo

Bárbara Malveira Orfanò[1]
Thaís Helena Pereira Marques[2]

RESUMO: Estudos baseados na interlíngua de aprendizes da língua inglesa têm ganhado destaque nos últimos anos. Este trabalho tem como objetivo analisar o discurso oral de um grupo de alunos da graduação do curso de Letras (Licenciatura em Inglês) da Universidade Federal de São João del-Rei, usando ferramentas de análise da Linguística de *Corpus* (palavras-chave, listas de frequência, linhas de concordância e pacotes lexicais). Alem disso propõe analisar os itens lexicais mais frequentes em um *corpus* de aprendizes, a fim de que se possam identificar aspectos importantes no discurso espontâneo dos alunos. Para tanto, foi compilado um *corpus* oral de pequena dimensão, coletado durante as aulas da disciplina "Habilidades Orais em Língua Inglesa", ministrada no primeiro semestre de 2011. O *corpus* de aprendizes foi contrastado com um *subcorpus* de falantes nativos *The Santa Barbara Corpus of Spoken Language*. Os resultados apontam diferenças interessantes tanto na forma quanto no uso dos pacotes lexicais em cada *corpus*.

PALAVRAS-CHAVE: *Corpus* de aprendizes, Frequência, Pacotes lexicais.

ABSTRACT: Studies based on the interlanguage of learners of English have gained prominence in recent years. This work aims to analyze the oral discourse of a group of graduate students from the Liberal Arts course (English) of the Federal University of Sao Joao del Rei, using analytical tools from Corpus Linguistics (keywords,

[1] Bárbara Malveira Orfano tem doutorado em Linguistica Aplicada: estudos baseados em corpora e atualmente é professor adjunta da Universidade Federal de São João del-Rei (UFSJ).

[2] Thais Helena Pereira Marques é graduanda do curso de Letras (Habilitação inglês) na Universidade Federal de São João del-Rei.

lists frequency, lines of lexical agreement and packages). In addition, it aims at analyzing the most frequent lexical items in a corpus of learners' oral discourse, so that one can identify important aspects in students' oral discourse. To that end, we compiled an oral corpus, collected in the discipline "Oral Skills in English," taught in the first semester of 2011. The learner corpus was contrasted with a subcorpus of native speakers *The Santa Barbara Corpus of Spoken Language*. The results show interesting differences both in the forms and uses of lexical bundles.

KEYWORDS: Learner corpus, Frequency, Lexical bundles

1 Introdução

Nos últimos anos, estudos baseados na Linguística de *Corpus* têm contribuído para a Linguística Aplicada ao possibilitar a descrição da linguagem oral ou escrita de aprendizes de uma língua. Mais especificamente, pesquisas baseadas em *corpora* de aprendizes têm ganhado notoriedade no campo das linguísticas descritiva e aplicada ao ensino. Biber, Conrad e Reppen (1998, p. 26) observam que muitos pesquisadores têm investigado a aquisição da segunda língua, compilando *corpora* escritos ou orais de aprendizes. Por meio desses estudos, são obtidos dados que podem ser utilizados para vários fins, como preparação de materiais didáticos, compilação de dicionários e descrição da interlíngua de aprendizes. Além disso, também são úteis a professores e alunos no que concerne à linguística de *corpus* aplicada ao ensino pela identificação das características do uso excessivo (sobreuso) de uma palavra ou expressão, definição de erros praticados por aprendizes e descrição das transferências de características da língua materna para a segunda língua, além de estabelecer distinções entre performances linguísticas de nativos e não-nativos (MEUNIER, 2010).

Entretanto, ressaltamos que a maioria dos estudos baseados em *corpora* são focados na habilidade de escrita. No que diz respeito a *corpora* orais, nota-se que pesquisas com foco nessa modalidade linguística ainda são bastante incipientes devido às dificuldades envolvendo compilação e transcrição dos dados.

O presente trabalho se insere dentro das pesquisas sobre *corpus* de aprendizes e tenta contribuir para o aumento de trabalhos em

torno da descrição da interlíngua oral de aprendizes. Para tanto, foi compilado um *corpus* oral de aprendizes graduandos em inglês do curso de Letras da Universidade Federal de São João del-Rei. Esse *corpus* foi contrastado com o *corpus* paralelo de nativos americanos, *The Santa Barbara Corpus of Spoken English*, para que fossem observadas as características linguísticas produzidas oralmente pelo grupo de aprendizes.

Os resultados obtidos serão explicitados na análise encontrada neste estudo. Anteriormente a ela, será feito um panorama cronológico dos principais *corpora* orais compilados e suas características para a melhor contextualização dos avanços da Linguística de *Corpus*. Por isso, chamamos a atenção para os *corpora* mais significativos, que vão desde 1959 até a presente data.

2 Referencial teórico

Um dos primeiros *corpora* a serem compilados foi o *Survey of Usage English* (SEU) em 1959. Esse *corpus* foi importante, pois, por meio dele, foram delimitados os critérios que caracterizam um *corpus*, como números definidos de textos e palavras, servindo como referência a *corpora* posteriores (SARDINHA, 2004).

Durante a década de 1960, foi organizado o primeiro *corpus* do gênero, chamado *Brown University Standard Corpus of Present Day American English*, contendo um milhão de palavras da língua inglesa escrita. Ele foi lançado em uma época na qual a coleta de dados linguísticos era vista com descrédito, uma vez que a vertente gerativista promovida pelo livro *Sintatic Structures,* de Noam Chomsky, se contrapunha a essa perspectiva. Ela se baseava na gramática gerativa, que sustentava a ideia de que pesquisas sobre uma língua não podem ser obtidas por meio de um *corpus*, pois o verdadeiro conhecimento provinha da intuição linguística de nativos de uma língua (BIBER *et al.*, 1998).

Com o avanço da tecnologia e o uso dos computadores em ambientes acadêmicos, a linguística de *corpus* se desenvolveu, facilitando a compilação de *corpora* eletrônicos e possibilitando também compilações cada vez maiores, chamados de *megacorpora* (SARDINHA, 2004).

Um exemplo desses *corpora* é o *British National Corpus* (BNC). Esse *corpus* consiste de mais de 100 milhões de palavras, no qual se

têm exemplos da língua inglesa moderna falada, por meio de conversas informais, e escrita, extraídos de jornais, periódicos, livros acadêmicos e revistas.

Outro *corpus* do gênero é o *Cambridge Nottingham Corpus of Discurse in English* (CANCODE), que é um *corpus* que contém cerca de cinco milhões de palavras, extraídas a partir da gravação de conversas espontâneas de nativos da língua inglesa britânica, por meio de vários tipos de interações sociais, como pedidos de informação, conversas entre amigos e discussões, dentre outras.

Há também outros *corpora* disponíveis *online*, como o *Corpus of Contemporary American English* (COCA) e o *Lexical Tutor* (LEXTUTOR). O COCA é composto por dados orais e escritos retirados de jornais, revistas, artigos acadêmicos, livros de ficção e conversas espontâneas, que somam mais de 425 milhões de palavras. Além dele, o LEXTUTOR é um concordanciador *online* desenvolvido por Thomas Cobb, da Universidade de Quebec, que possibilita a análise de palavras isoladas e de seus colocados por qualquer usuário, podendo ser acessado pelo endereço eletrônico http://www.lextutor.ca/.

Nos últimos anos, a Linguística de *Corpus* tem contribuído com pesquisas nos domínios da Linguística, como na descrição das línguas, da Linguística Aplicada, relacionada ao ensino e aprendizagem de línguas, e da Linguística Computacional, no desenvolvimento de *softwares* específicos (ALUÍSIO; ALMEIDA, 2006), entre outras áreas de conhecimento. O presente trabalho posiciona-se dentro da Linguística Aplicada, tendo como foco principal a produção oral de aprendizes de língua inglesa para a compilação de um *corpus*. Sendo assim, esse referencial teórico incluirá também o histórico dos *corpora* de aprendizes importantes, seguido de estudos que servem de base para esta pesquisa.

Para a observação de tais características produzidas por aprendizes, algumas universidades compilaram seus próprios *corpora*, que contribuíram para o avanço dos estudos baseados em Linguística de *Corpus*. Podemos citar exemplos como os *corpora Louvain Corpus of Native Essays* (LOCNESS), *International Corpus of Learner English* (ICLE) e seu *subcorpus Brazilian International Corpus of Learner English* (Br-ICLE).

O LOCNESS também é um *corpus* monitor, que contém 324.304 palavras (SHEPHERD, 2009, p. 106) extraídas de redações acadêmicas de nativos americanos e britânicos, das quais 60.209 palavras são provenientes de redações de alunos britânicos do período inicial da universidade, 95.695 palavras de estudantes britânicos de outros

períodos e 168.400 palavras resultantes de redações de universitários americanos em geral.

Outro *corpus* de aprendizes representativo do gênero é o *corpus* desenvolvido pelo projeto ICLE, também compilado pela Universidade Católica de Louvain, que é composto por redações em língua inglesa de aprendizes universitários de níveis intermediário e avançado de 16 países, o qual coletou mais de 3,7 milhões de palavras (DUTRA; SILERO, 2010, p. 917). Esse *corpus* tem sido usado para pesquisas que analisam as características lexicais, gramaticais e discursivas produzidas por aprendizes (GILQÜIN; GRANGER, 2011). O Brasil está inserido nesse projeto, representado pelo *subcorpus* Br-ICLE. Esse *subcorpus* foi coletado a partir de redações de alunos universitários brasileiros de níveis avançados de inglês, que estão cursando do quinto período em diante na universidade.

A análise da produção de aprendizes tem ganhado importância no meio acadêmico. Gilqüin e Granger (2011) compararam o uso da preposição *into* utilizada no *corpus* ICLE por aprendizes holandeses, franceses, espanhóis e tswanos. De acordo com os resultados obtidos, as pesquisadoras perceberam que os holandeses estão mais próximos da frequência de uso do *corpus* de nativos britânicos, seguidos dos franceses e espanhóis. Por essa análise, Gilqüin e Granger (2011) observam que o tipo de exposição à língua é refletido pelo uso de tais expressões.

Dutra e Sardinha (2010) observaram a produção escrita de aprendizes brasileiros por meio do *corpus* Br-ICLE e concluiu que esses alunos fazem sobreuso de auxiliares, além de se utilizarem do pacote lexical *I think* com frequência. Sardinha conclui que os aprendizes fazem uso da linguagem oral na escrita e que as redações são semelhantes ao gênero de escrita escolar.

Dutra e Silero (2010) utilizaram-se de textos argumentativos de alunos universitários obtidos por meio do *Corpus* de Aprendizes Brasileiros do Inglês (CABrl), para pesquisar a ocorrência de pacotes lexicais que contêm as preposições *for* e *to*, comparando-os com os *corpora* Br-ICLE e LOCNESS. A partir dos dados encontrados, elas concluíram que alguns pacotes lexicais, como *prepare for, wait for, search for* e *live for*, são igualmente usados por nativos. Por outro lado, as autoras perceberam que houve sobreuso da expressão *consider for* e uso inadequado das expressões *contribute to* e *spread to*, que foram utilizadas com a palavra *for* pelos aprendizes. A partir dessas observações, Dutra e Silero (2010) sugerem atividades

pedagógicas com uso de linhas de concordância, para que os alunos percebam o uso correto de tais expressões.

Recentemente, alguns pesquisadores têm centrado seus estudos na fraseologia. A fraseologia baseia-se no estudo da frase pelo fato de ela ser uma unidade primária de significado (DUTRA; SARDINHA, 2010). Por meio das relações entre unidades de palavras, as quais vão sendo aprendidas e cristalizadas durante o uso pela sociedade, formam-se os pacotes lexicais (chamados também de *n-grams*, *clusters, chunks*, unidades múltiplas de palavras, dependendo dos objetivos de cada pesquisa). Os pacotes lexicais são palavras que frequentemente co-ocorrem em uma língua e são divididos entre *collocations* (colocações), *phrasal verbs* (verbos frasais) e *idioms* (expressões idiomáticas), e a frequência deles pode ser analisada pela linguística de *corpus* (GRANGER; MEUNIER, 2008; DUTRA; SARDINHA 2010).

A linguística de *corpus*, portanto, alia-se à fraseologia para analisar estatisticamente a ocorrência desses pacotes lexicais. Por meio dessas análises, são observadas as frequências de uso desses pacotes, o que pode ser útil a aprendizes de línguas.

Dutra e Sardinha (2010) analisaram os pacotes lexicais extraídos dos *corpora* LOCNESS e ICLE e os contrastaram com os pacotes lexicais do *subcorpus* de aprendizes brasileiros Br-ICLE. A partir dessa análise entre os *corpora*, os pesquisadores perceberam que as expressões mais amplamente usadas nas redações são as referenciais, como no exemplo mais frequente encontrado *The fact that*. Dutra e Sardinha (2010) também observaram que no *corpus* Br- ICLE há pouca variedade de uso dos pacotes lexicais e, além disso, os usos desses pacotes são diferentes, principalmente se comparados com o *corpus* ICLE.

Baseando-se em Lewis (1996, 2000) e Richards (1994), Matte (2009, p. 56) observou o processo de aquisição do vocabulário linguístico de um grupo de alunos e concluiu que o uso de pacotes lexicais propicia um melhor aprendizado. Ela exemplifica essa perspectiva de ensino, apresentando o exemplo de um aluno, que pergunta ao professor o significado da expressão *game over* que estava em um texto. De acordo com a autora, os professores poderiam ampliar o vocabulário dos alunos a partir de expressões como essa, explorando o significado da palavra *game*, utilizando-se de sinônimos como *The game finished*, ou motivando-os a aprender pacotes lexicais como *Can I play a game?*

Matte (2009) acredita que o ensino deve envolver a associação de palavras aos contextos em que elas estão inseridas, pois, assim,

os alunos terão mais chances de retenção desse vocabulário na memória. Por meio das aulas ministradas na universidade, verificou-se que o principal motivo para que os alunos desse grupo ansiassem por aulas de conversação foi por que muitos deles almejavam alcançar a fluência na língua inglesa. Ellis (2008) observa que o uso de expressões fixas otimizam o alcance da fluência, pois esses itens são bastante frequentes durante a linguagem oral. Por isso, é importante que estudos em torno da fraseologia sejam explicitados nesta pesquisa.

3 Participantes e metodologia

A presente pesquisa teve início a partir da necessidade de melhora das competências orais, observadas pelos próprios graduandos da Universidade Federal de São João del-Rei (UFSJ).[3] Atendendo a esse pedido, no primeiro semestre de 2011, foi oferecida a disciplina "Habilidades Orais em Língua Inglesa", com carga horária de 60 horas. Essas aulas foram gravadas, transcritas e analisadas, e compõem o *corpus* de aprendizes desta pesquisa. No total, o *corpus* resultante contém cerca de 25 mil palavras contendo a produção oral dos alunos matriculados nessa disciplina.

Ao todo, participaram 18 alunos, cinco homens e 13 mulheres, com idade média de 23 anos, todos graduandos em inglês, com variados níveis de proficiência nessa língua.

Durante as aulas, foram discutidos vários assuntos, que abordaram desde a vida cotidiana, como o "dar um tempo" em um namoro, agressões físicas sofridas por mulheres e fatos da infância dos próprios alunos, até acontecimentos que foram notícia durante o período, como a vinda de Barack Obama ao Brasil e a chacina ocorrida em uma escola do estado do Rio de Janeiro no ano de 2011.

Serão descritas, a seguir, a estrutura e a metodologia empregadas para uma melhor ilustração das aulas ministradas, partindo de um exemplo utilizado em uma das aulas. Primeiramente, o tema a ser abordado foi apresentado por meio de um vídeo, no qual a apresentadora americana Oprah Winfrey visita o cantor Michael Jackson. A partir dele, pequenos grupos foram formados, para que

[3] Os alunos do 5º período do curso de licenciatura em inglês fizeram um abaixo assinado no segundo semestre de 2010, solicitando o oferecimento de uma disciplina que focasse nas habilidades orais de língua inglesa.

fosse feito um exercício escrito, que continha algumas perguntas relativas ao vídeo, como as expostas no Anexo I deste trabalho.

Além disso, apresentações orais sobre temas escolhidos pelos próprios alunos também fazem parte do *corpus*. A maioria dos temas abordados foi de cunho biográfico sobre personalidades nacionais e estrangeiras do cinema, da literatura, da política e da música, com as quais os alunos se identificaram, como Jô Soares, Michael Jackson, Dilma Rousseff, Tim Maia, Margareth Atwood, Tim Burton, John Bon Jovi, Walt Whitman, Maria Carrey e Grace Kelly. Também houve duas apresentações de alunos, nas quais as histórias de uma avó e de uma tia foram apresentadas.

Posteriormente, as gravações das aulas e das apresentações foram recolhidas e transcritas, utilizando o *Express Scribe,* que é um *software* gratuito para transcrições de arquivos de áudio, o qual pode ser baixado gratuitamente pelo endereço eletrônico http://www.nch.com.au/scribe/index.html para uso de ferramentas simples. Se preferível, é possível que esse *software* seja comprado no mesmo *site*, para trabalhos que necessitem de recursos mais completos. Além disso, para a uniformização do *corpus* de aprendizes, foram aplicadas as convenções de transcrição utilizadas pelo *corpus* oral de aprendizes *LINDSEI* (*Louvain International Database of SpokenEnglish Interlanguage)* da Universidade de Louvain, Bélgica.[4]

O *corpus* de nativos usado para contraste nesta pesquisa é o *Santa Barbara Corpus of Spoken Language,* que contém cerca de 200 mil palavras com diálogos de 30 minutos, os quais descrevem aspectos da produção oral de falantes nativos do inglês americano espontâneo, contendo várias idades, etnias, gêneros, posições sociais, origens e regiões do país. Esse *corpus* contém uma variedade de contextos, incluindo conversas espontâneas no ambiente familiar, diálogos entre amigos no *campus* da universidade ou entre colegas de trabalho, conversas telefônicas, rituais religiosos e pessoas em momentos de descontração, como em um jogo de cartas ou no momento de uma estória, dentre outros. No entanto, para esta pesquisa, foi selecionado um *subcorpus* de 25 mil palavras, a fim de que a análise entre os *corpora* seja precisa tanto quantitativa quanto qualitativamente.

Para a análise dos dados obtidos e contraste entre os *corpora*, foi utilizado o *software Wordsmith Tools versão 3.0.* Esse é um programa

[4] http://www.uclouvain.be/en—cecl—lindsei.html.

para descrição linguística com *corpus* (SARDINHA, 2004, p. 83), que pode ser adquirido pelo *site* http://www.lexically.net/wordsmith/, para demonstração, com ferramentas limitadas, ou pela compra da licença para uso pleno dos recursos dele.

No entanto, para análise dos dados deste trabalho, foram utilizadas somente as ferramentas básicas desse programa, como a lista de palavras, *n-grams* e linhas de concordância.

4 Análise

A partir dos dois *corpora*, selecionamos os dados mais recorrentes em ambos e analisamos os pacotes lexicais contendo duas, três e quatro palavras. Os dados encontrados nos dois *corpora* estão listados no quadro comparativo a seguir:

QUADRO 1
Comparação entre os corpora *Corpus* Oral UFSJ
e *Santa Barbara Corpus of Spoken*

Corpus Oral UFSJ			*Santa Barbara Corpus of Spoken English*		
N	Pacote Lexical	Freq.	N	Pacote Lexical	Freq.
1	**I THINK**	124	1	**YOU KNOW**	122
2	**I DON'T KNOW**	70	2	AND I	62
3	**YOU KNOW**	57	3	AND THEN	50
4	**HOW CAN I SAY**	43	4	I WAS	50
5	KIND OF	39	5	I MEAN	45
6	I THINK THAT	28	6	IT WAS	45
7	DO YOU THINK	17	7	**I THINK**	40
8	I DON'T THINK	17	8	AND THEY	35
9	AND SO ON	13	9	**I DON'T KNOW**	30
10	I THINK IT'S	12	10	AND I SAID	11

Após uma análise preliminar desse quadro, podemos inferir que existem diferenças de formas e de frequência quanto ao uso dos pacotes lexicais nos dois *corpora*. Essas diferenças e suas implicações para o discurso, principalmente do *corpus* de aprendizes, estão no cerne deste trabalho. Levando em consideração os objetivos principais da pesquisa, analisaremos os itens mais frequentes no *corpus* de aprendiz, contrastando-os com o *corpus* de falantes nativos.

No entanto, analisaremos somente os pacotes lexicais *I think, I don't know, you know* e *how can I say* encontrados como mais recorrentes no *corpus* principal. Da mesma forma, serão usados os mesmos pacotes lexicais contidos no *corpus* paralelo. Os pacotes lexicais serão analisados separadamente e contrastados com o *corpus* de nativos americanos no decorrer desta seção.

4.1 I Think

Ao analisar o verbo *to think*, Biber *et al.* (1999) e Granger e Parquot (2009) observam que ele é marcado como característico de linguagem falada. Como um verbo lexical, o item faz parte da categoria dos verbos mentais e emotivos (GRANGER; PARQUOT, 2009). Além disso, pesquisas anteriores já demonstraram que o pacote lexical *I think* é mais frequentemente usado por aprendizes em geral, e não só por brasileiros (SARDINHA, 2010).

De acordo com os resultados da pesquisa, o pacote lexical *I think* foi o mais frequente, usado 124 vezes no *corpus* de aprendizes. No entanto, ao analisarmos a frequência desse pacote lexical no *corpus* paralelo, podemos perceber que ele não é tão recorrente, pois foi utilizado somente 40 vezes pelos falantes nativos, ocupando a sétima colocação no quadro comparativo.

Para ilustrarmos nossa análise, apresentamos algumas amostras retiradas do *corpus* paralelo e do *corpus* de aprendizes por meio dos excertos a seguir.

Excerto 1

[*Corpus Santa Barbara*: amigas conversando sobre um livro que ambas estão lendo.]
<DARRYL> But, but to try and and talk me out of believing in Murphy's Law, by offering a miracle as a replacement, that doesn't d- work.
<PAMELA> Well you're right, **I think** they're probably flip sides.

Excerto 2

[*Corpus* de aprendizes: fragmento de uma das aulas, na qual foi discutida a chacina ocorrida no Rio de Janeiro em 2011.]
<C> if you were a a par = a mother a father . what would you think about it </C>
<D> ... **I think** . (eh) they need . blame someone <starts laughing> .. they will blame the guy <XX> and will blame the teacher that let him in they will blame the school or the government about this <X> humans **... I think** I . if I was mother I . would blame the adults it's not (mm) </D>

A partir da análise das amostras dos *corpora*, concluímos que o pacote lexical *I think* é utilizado pelos aprendizes como uma tentativa de diminuição da assertividade, o que é um indício de que o grupo de estudantes presta mais atenção a traços de solidariedade do que o grupo de nativos (GOFFMAN, 1967; BROWN, LEVINSON, 1987).

4.2 I don't know

Ao analisarmos esse pacote lexical no *corpus* principal, verificamos que ele é sobreusado pelos aprendizes do grupo, sendo utilizado aproximadamente duas vezes mais que o grupo do *corpus* paralelo. A seguir, demonstramos, por meio das linhas de concordância extraídas do programa *Wordsmith Tools*, como esse pacote lexical foi utilizado pelos aprendizes.

N	concordance	
1	ves they can they kill themselves .	**I don't know** why it's it's it seem
2	l he has shots . (eh) in the in two	**I don't know** if it was children or
3	at has happened to him . or maybe .	**I don't know** I think that a person
4	.. so so only . in united states ..	**I don't know** if they . that I . I
5	the guy . (em) killed himself no no	**I don't know** .. actually is is is
6	s story because I was working (hmm)	**I don't know** about the story When
7	Let's talk yes I	**I don't know** nothing about this sto
8	se it's common things like that and	**I don't know** it was a it was a at
9	nd of arsenal and inside the this .	**I don't know** . his bala and his su
10	s a extra (em) bullets lots of (eh)	**I don't know** how many bullets bala
11	mmon it's just maybe this guy had a	**I don't know** some mental problems
12	r very . but kill a lot of kids and	**I don't know** if . there . they lik
13	nother school for the children or .	**I don't know** what's going to happe
14	this term was discussed now . (eh)	**I don't know** not now now now but i
15	somehow rejected by the girls . or	**I don't know** it could be also an e
16	be it was cultural and then injured	**I don't know** if it's cultural I th
17	joking but . they have on screams .	**I don't know** . I've rom kill someo
18	politician . a (mm) politician ...	**I don't know** (eh) someone .. a . i
19	errorisma in a in a Iraquian school	**I don't know** I don't remember very
20	. something like a ... te terroris	**I don't know** how to say terrorism

Orfanò (2010) observa que o pacote lexical *I don't know* é utilizado quando o falante não tem certeza do que foi dito por ele ou quando há intenção de menos assertividade durante uma conversa. No *corpus* de aprendizes, podemos perceber que os exemplos encontrados no *corpora* são consonantes com os dados da pesquisa conduzida por Orfanò (2010), pois os aprendizes utilizaram-se desse pacote lexical tanto para demonstrar desconhecimento do assunto abordado em sala de aula quanto para expressarem falta de assertividade.

4.3 You know

Observando o quadro comparativo, podemos perceber que há grande diferença na frequência de uso desse pacote lexical. No *corpus* de nativos, essa expressão se encontra em primeiro lugar, usada 122 vezes, enquanto no *corpus* de aprendizes, o pacote lexical se encontra em terceiro lugar, usado 57 vezes. Isso denota que o termo *you know* é subusado pelo grupo de aprendizes. Pesquisas anteriores já analisaram o pacote lexical *you know*. Portanto, essa expressão pode ser considerada como recorrente não só no *corpus* da pesquisa como em vários outros *corpora*.

Segundo McCarthy e Carter (2002) e Orfanò (2010, p. 129), o pacote lexical *you know* é um marcador discursivo que denota compartilhamento de conhecimento entre pessoas em um diálogo. No que se refere ao *corpus* de aprendizes, concluímos que os estudantes utilizaram-se dessa expressão não só para compartilharem conhecimentos, mas também para terem a confirmação das ideias que foram ditas por eles. Isso pode ser percebido por meio do excerto a seguir.

Excerto 3

[*Corpus* de aprendizes: fragmento de uma das aulas, na qual o tema foi "dar um tempo" em um namoro.]
<F> you have a good lucky because my my ex are not accepting me anymore and even if I want him he doesn't because I don't know because he's very he's a hard person <XXX> <overlap> </G>
<D> <XXX> with his family </D>
<F> he don't accept me not <X> </F>
<D> why not </D>
<F> I don't know it's a </F>
<D> he is a hard person </D>
<F> he is a hard person and it's difficult to understand people's mind **you know** </F>
<E> so so that's why you can <overlap> how can I say </E>
<D> <XXX> (mm) just you </D>
<F> I tried to: to ask </F>

Porém, percebemos que no *corpus* de nativos a expressão *you know* é comumente utilizada na função de compartilhamento de ideias. Sendo assim, podemos inferir que, enquanto os falantes nativos utilizam-se de pacotes lexicais dentro da função de compartilhamento de ideias, os aprendizes tendem a preferir o pacote

lexical *you know* para confirmar ideias e/ou opiniões.

Excerto 4

[*Corpus* de nativos: fragmento de uma conversa sobre um jantar.]
<MARILYN> But I **you know**, I said, I want it to be homemade.**You know**, something special.
<PETE> Well at [least] they are uh like already breaded —

4.4 How can I say

Ao comparar a recorrência desse pacote lexical entre os *corpora*, podemos perceber que ele não está presente no *corpus* de *Santa Barbara*. Porém, ele é a quarta expressão mais recorrente do *corpus* da pesquisa, produzida 43 vezes pelos aprendizes. Para melhor análise dos dados, optamos primeiramente por demonstrar, a seguir, as expressões utilizadas pelo grupo de aprendizes da pesquisa por meio das linhas de concordância extraídas do *wordsmith tools*.

N	Concordance	
1	he (eh) she (eh) 'cause he was	**how can I say adotado** adopt .
2	(em) the . the media come (em)	**how can I say** shot he **filmou**
3	ike you can't you may use (eh)	**how can I say** you must use gl
4	I say a participating **palestra**	**how can I say** a speech he mad
5	n children and teenagers and .	**how can I say** he hurts twent
6	won don't want we .. (em) take	**how can I say férias** it's .
7	d in the letter he ask to be .	**how can I say enterrado** buried
8	portant for us to . to .. (em)	**how can I say** regard to to th
9	hocking then they don't even .	**how can I say** . think . (em)
10	y . is something that (eh) ...	**how can I say foge do controle**
11	right some surrealistics (eh)	**how can I say** not only inspire
12	s is dangeour when someone you	**how can I say ressentimento**
13	know so so that's why you can	**how can I say** (mm) just you I
14	idn't accept I I guess he (mm)	**how can I say** he has some (mm)
15	ke that yeah will cheat on and	**how can I say trair** in english
16	ke that condition don't want go	**how can I say** hug to him yeah
17	eak have a breaking up .. (mm)	**how can I say** (mm) light brea
18	use you see that relation have	**how can I say** not fighting ar
19	w I don't agree with this this	**how can I say** (mm) this what
20	elly this to find a a point to	**how can I say equilíbrio** bala

Considerando essas linhas de concordância, verificamos que, geralmente, o pacote lexical *How can I say* tem como colocado, uma palavra em português, seja anterior ou posterior à frase. Após manualmente pesquisar as linhas de concordância, verificamos que a expressão ocorre 44% das vezes, tendo como colocado uma palavra em português.

A partir dessa observação, concluímos que os alunos utilizam essa expressão quando não têm ao seu dispor vocabulário específico relacionado ao tema discutido pelos alunos durante a aula em questão. É importante ressaltar também que essa frase é comumente ensinada em aulas de língua inglesa. Consequentemente, essa expressão se torna uma "marca registrada" do discurso oral do aprendiz.

Não há ocorrência dessa expressão no *corpus* paralelo utilizado para esta pesquisa. Conclui-se que esse item linguístico é praticamente inválido para o falante nativo. Como mencionado em Oliveira e Orfanò (2011), alguns pacotes lexicais pertencem exclusivamente ao discurso do aprendiz devido ao tipo de instrução recebida pelos alunos durante a experiência de aprendizado em língua inglesa.

Entretanto, nota-se que em alguns exemplos o uso da expressão *How can I say* não são utilizados na forma já descrita. Podemos perceber que, algumas vezes, esse pacote lexical foi usado como uma forma de "pensar alto", enquanto procuram a palavra adequada, por exemplo, *hocking then they don't even . **how can I say** . think . (em) like many have had.*

Concluímos que o pacote lexical *how can I say* dentro da função de (pensar alto e/ou procurar o significado de uma palavra) foi utilizado especificamente por alunos do grupo que são mais proficientes na língua inglesa.

5 Conclusão

Os resultados apresentados neste trabalho evidenciam que as diferenças de forma e frequência dos pacotes lexicais utilizados pelos participantes caracterizam de maneira específica cada grupo. Analisando o item *I think*, percebemos que esse pacote lexical foi utilizado três vezes mais no *corpus* de aprendiz do que no de falantes nativos. Tal dado demonstra que o grupo de aprendizes observado

nesta pesquisa demonstra maior preocupação com traços de solidariedade e assertividade, diferentemente do *corpus* de nativos utilizados no presente trabalho.

Além disso, foi verificado que o pacote lexical *I don't know* é usado de duas maneiras: para demonstrar falta de conhecimento sobre o assunto abordado e também para expressar falta de assertividade.

Os aprendizes do grupo utilizaram-se do pacote lexical *you know* como uma forma de confirmarem suas ideias por meio do discurso. Algumas vezes, esse pacote pode ser considerado como a versão reduzida do pacote lexical maior *you know what I mean*.

A expressão *How can I say* é recorrente no discurso dos aprendizes. Por outro lado, ela não foi encontrada no *corpus* dos nativos. Isso significa que *How can I say* é própria da fala do grupo de aprendizes. Por meio das análises, concluímos que o grupo de aprendizes faz uso dos pacotes lexicais discutidos na seção dedicada à análise dos dados para seguirem normas de polidez, enquanto os estudantes nativos do *corpus* de *Santa Barbara* tendem a não se preocupar necessariamente com tais normas. Analisar os pacotes lexicais à luz das teorias de polidez ultrapassa os objetivos do presente trabalho. Entretanto, os resultados aqui observados servem como base para futuras pesquisas que em muito contribuirão para a compreensão da interlíngua oral de aprendizes.

Referências Bibliográficas

ALUÍSIO, Sandra Maria; ALMEIDA, Gladis Maria de Barcellos. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio*, v. 4, n. 3, p. 155-177, set./dez.2006.

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

BIBER, Douglas; CONRAD Susan; REPPEN, Randi. *Corpus linguistics-Investigating language structure and use.* Cambridge: Cambridge University Press, 1998.

BROWN, P.; S. C. LEVINSON. *Politeness: some universals in language use*. Cambridge: Cambridge University Press, 1987.

DUTRA, D. P; SILERO, Rejane Protzner. Descobertas linguísticas para pesquisadores e aprendizes: a Linguística de *corpus* e o ensino de gramática. *RBLA*, Belo Horizonte, v. 10, n. 4, 2010.

DUTRA,D.P; BERBER SARDINHA, Tony. *Pacotes lexicais em corpora de aprendizes*, Trabalho apresentado no ELC 2010, Porto Alegre.

ELLIS, N. C. Phraseology: The periphery and the heart of language. In. MEUNIER, F.; GRANGER, S. A. (Eds.). *Phraseology in foreign Language Learning and Teaching*. Amsterdam: John Benjamins, 2008. p. 1-13.

GILQÜIN, Gaëtanelle; GRANGER Sylviane. *From EFL to ESL*: Evidence from the International Corpus of Learner English. 2011.

GOFFMAN, E. *Interaction Ritual. : essays on face-to-face behaviour*. New York: Anchor Books, 1967.

GRANGER, Sylviane; PAQUOT, Magali. Lexical verbs in academic discourse: a corpus-driven study of learner use. In: CHARLES, M.; PECORARI, D. HUNSTON, S. (Ed.). *Academic Writing: At the Interface of Corpus and Discourse*. London/New-York: Continuum, 2009. p. 193-214.

GRANGER, Sylviane; MEUNIER, Fanny. *Phraseology in foreign language learning and teaching*. John Benjamim Publishing Company, Amsterdam, 2008.

MATTE, Fátima. Como os chunks facilitam o aprendizado do vocabulário da língua inglesa. *Signos*, ano 30, n. 2, p. 55-64, 2009.

McCARTHY, M.J.; CARTER, R. From conversation to corpus: a dual analysis of a broadcast political interview. In: SANCHEZ-MACARRO, A. (Ed). *Windows on the World*: Media Discourse in English. Valencia: University of Valencia Press, 2002. p 15-39.

MEUNIER, Fanny. Learner Corpora and English Language Teaching: Check up time. *International Journal of English Studies*, v. 21, n. 1, p. 209-220, 2010.

OLIVEIRA, A.; ORFANÒ, B. *How Brazilian learners express modality in their writing: a corpus-based study of lexical bundles. Caderno de resumos do 18 INPLA*. São Paulo: PUC-SP. p. 78. 2011.

ORFANO, B. The Representation of Spoken Language: a corpus-based study of sitcom discourse. 2010. 305 f. Tese (Doutorado) - Mary Immaculate College - University of Limerick, 2010.

SHEPHERD, Tania M. G. *Corpora* de aprendiz de língua estrangeira: um estudo contrastivo de n-gramas. *Veredas Online – Linguística de Corpus e Computacional*, Juiz de Fora: PPF Linguística/UFJF, v. 2, p. 100-116, 2009.

Sites Consultados

Br- ICLE <http://www2.lael.pucsp.br/corpora/bricle/projeto.htm> (acesso em 24-11-2011)

BNC <http://www.natcorp.ox.ac.uk/corpus/index.xml> (acesso em 04-12-2011)

CANCODE <http://www.cambridge.org/pl/elt/catalogue/subject/custom/item3646595/Cambridge-English-Corpus-Cambridge-and-Nottingham-Corpus-of-Discourse-in-English-(CANCODE)/?site_locale=pl_PL> (acesso em 10-01-2012)

Express Scribe <http://www.nch.com.au/scribe/index.html> (acesso em 24-11-2011)

LOCNESS <http://www.uclouvain.be/en-cecl-locness.html> (acesso em 10-01-2012)

Santa Barbara corpus of Spoken English <http://www.linguistics.ucsb.edu/research/sbcorpus.html> (acesso em 06-12-2011)

Anexo I

Fragmento de uma das aulas ministradas

You will hear one of the most famous interviews ever made on television.
Read the questions and answer them:

PART I
1- Where does he live?
2- What rooms does she mention in her interview?
3- How many siblings did he have?
4- Who was getting most of the attention on Jackson's Five?
5- When they were deciding about the interview, what was their agreement?
6- Was he nervous during the interview?
7- How did he describe the time he used to sing with his brothers?
8- What does he think about James Brown?
9- How did he feel when he was on stage?
10- How did he feel when he was off-stage?

PART II
Answer true or false for the following statements.
a) ( ) He did not believe that he lost his childhood.
b) ( ) Michael Jackson was always around kids to compensate for his past.
c) ( ) He agreed on what Latoya said about his family.
d) ( ) He somehow felt angry with his father.
e) ( ) His mother was also very strict as a parent.