IX Escola Brasileira de Linguística Computacional

EBRALC 2017

XIV Encontro de Linguística de Corpus

ELC 2017

Organização:

Apoio:

Editora da Universidade do Vale do Rio dos Sinos
Editora Unisinos
Av. Unisinos, 950
93022-000 São Leopoldo-RS-Brasil
editora@unisinos.br
www.edunisinos.com.br

Caderno de Resumos ELC-EBRALC 2017 corresponde à 1ª edição em versão eletrônica. O **Caderno de Resumos ELC-EBRALC 2017** é uma publicação em que estão compiladas versões recortadas pelos organizadores de resumos das palestras, dos minicursos, dos *workshops* e dos textos dos resumos dos trabalhos (português e inglês) aprovados para apresentação na IX Escola Brasileira de Linguística Computacional (EBRALC 2017) e no XIV Encontro de Linguística de Corpus (ELC 2017). Os resumos, assim apresentados, estão organizados por modalidade de apresentação se seguem formatos diferentes, de acordo com os tipos de trabalhos envolvidos. Apenas o trabalho premiado no evento – *paper* 09, modalidade *short paper* – categoria Mérito Científico, está reproduzido na íntegra.

As informações deste Caderno de Resumos conjugam-se às publicadas no *site* do evento: http://www.ufrgs.br/elc-ebralc2017

ISBN: 978-85-98396-11-8

Revisão: Carlos Batanoli Hallberg
Editoração eletrônica: Fenando Piccinini Schmitt

Organização
Aline Evers
Aline Nardes
Larissa Brangel
Maria José Bocorny Finatto
Rove Luiza de Oliveira Chishman

Agradecimento especial
PPG-Letras UFRGS
Universidade Federal do Rio Grande do Sul
PPGLA-UNISINOS
Universidade do Vale do Rio dos Sinos

Editora da Universidade do Vale do Rio dos Sinos
Editora Unisinos
Av. Unisinos, 950
93022-000 São Leopoldo-RS-Brasil
editora@unisinos.br
www.edunisinos.com.br

Apoios – editais de órgãos de fomento à pesquisa:

CAPES FAPERGS CNPq

APRESENTAÇÃO

Aline Evers
Aline Nardes
Larissa Brangel
Maria José Bocorny Finatto
Rove Luiza de Oliveira Chishman

Nesta publicação, estão apresentados os resumos simples e resumos expandidos dos *workshops*, minicursos e minicursos expressos, das Mesas Redondas e dos trabalhos submetidos e apresentados nas sessões Oral, Pôster e Trabalhos em Andamento (sessão "Minuto de Loucura") durante o nosso evento **ELC-EBRALC 2017**. Também tivemos, pela primeira vez, uma atividade prévia ao evento, o "Esquenta ELC", com uma palestra especial a cargo do nosso convidado internacional. Mais detalhes sobre o evento podem ser encontrados no nosso *site* (http://www.ufrgs.br/elc-ebralc2017).

O **XIV Encontro de Linguística de Corpus (ELC)** e a **IX Escola Brasileira de Linguística Computacional (EBRALC)**, dois eventos associados, ocorreram entre os dias 15 e 18 de agosto de 2017 em São Leopoldo/RS, na Universidade do Vale do Rio dos Sinos (Unisinos). Para que esse evento "duplo" fosse possível, tivemos a cooperação da Universidade Federal do Rio Grande do Sul (UFRGS) e da Unisinos, em meio a um convênio entre o PPGLA Unisinos e o PPG-Letras UFRGS.

Os resumos aqui apresentados aparecem em diferentes formatos, estabelecidos por seus autores, mas foram por nós livremente recortados, a partir dos textos submetidos, com a intenção de destacar seus pontos principais e de guiar, inicialmente, um acesso rápido aos interessados que participaram do evento. A heterogeneidade formal dos textos aqui colocados dá-se em função dos trabalhos e das participações pertencerem a categorias diferentes, cada um com um padrão de submissão ou de apresentação. Apenas o trabalho da modalidade Oral, número 09, *Towards Machine Translation from Brazilian Portuguese-to-Libras: A Corpus-Based, Morphosyntactic Analysis,* de Francisco Aulísio dos Santos Paiva, Plínio Almeida Barbosa, Pablo Faria e José Mario De Martino, premiado como **Mérito Científico** do evento, aparece na íntegra conforme submetido. Em breve, uma seleção especial desses trabalhos – todos no formato de artigos longos – será apresentada em livro associado ao nosso evento, lançamento que você poderá conferir também no nosso *site*.

Em 2017, a temática do nosso evento envolveu reflexões sobre as **Contribuições da Linguística de Corpus para os Estudos sobre Gêneros Textuais e Discursivos**. Em edições anteriores do ELC-EBRALC, os participantes já haviam discutido temas como **Convergência tecnológica para o processamento e análise das línguas** (2012), **Corpus, tecnologia e cultura** (2014) e **Contribuições da Linguística de Corpus para a pesquisa linguística e para os Estudos da Tradução** (2015). Em 2017, seguimos tratando também da interface entre Ciência da Computação – pelo Processamento da Linguagem Natural (PLN)/Linguística Computacional – e Estudos da Linguagem – com a Linguística de Corpus e diferentes áreas da Linguística.

Nesta edição, contamos com o apoio dos órgãos de fomento Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (**CAPES**), Conselho Nacional de Desenvolvimento Científico e Tecnológico (**CNPq**) e

Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (**FAPERGS**), pelo que muito agradecemos. Nossos demais apoiadores foram: a Embaixada do Governo Norte-Americano (Brasília/DF), o *Sketch Engine*, a Sociedade Brasileira de Computação (**SBC**), o **ETS-TOEFL**, o programa Idioma Sem Fronteiras (**ISF – UFRGS**). Registramos também a parceria da Universidade Federal de Ciências da Saúde de Porto Alegre (**UFCSPA**), da Pontifícia Universidade Católica do Rio Grande do Sul (**PUC-RS**) e da Universidade de Caxias do Sul (**UCS**), cujos docentes nos apoiaram em diferentes etapas da organização do evento. Também tivemos a parceria da Direção do **Instituto de Letras da UFRGS**.

Além dos apoios institucionais, estudantes de graduação e de pós-graduação da UFRGS e da Unisinos prestaram-nos auxílio cujo valor não se conseguiria medir, tendo atuado como monitores, auxiliares de secretaria e membros da comissão organizadora executiva.

Neste **Caderno de Resumos** estão relatos de pesquisas em Linguística de Corpus e em Linguística Computacional relacionados à Linguística Histórica, Linguística Aplicada, ao Ensino de Línguas e de Tradução, à Linguística Cognitiva, às Ciências da Informação, dentre outras. Nele você encontra uma síntese das reflexões sobre as contribuições da **Linguística de Corpus** para os Estudos sobre Gêneros Textuais e Discursivos a partir de análises realizadas com apoio da tecnologia. Esses trabalhos buscaram demonstrar e validar fatos e suposições sobre o funcionamento da língua em diferentes registros e cenários comunicativos mediados pelo texto, principalmente o texto escrito, mas também com abordagens sobre *corpora* de fala.

A diversidade dos trabalhos aqui posta, esperamos, espelha o quanto o "território" da **Linguística de Corpus** tornou-se um importante ponto de encontro, um campo aberto para a cooperação e o diálogo respeitoso e crítico entre pessoas que, de algum modo e cada uma do seu jeito, se ocupam dos fenômenos da linguagem humana.

São Leopoldo-RS, em outubro de 2017.

****O número indicado acima do título do trabalho corresponde ao número de inscrição/submissão publicado no *site* do evento.**

SUMÁRIO

Resumos dos minicursos, minicursos expressos e *workshops* especiais

Workshop convidado

Palestra de encerramento do ELC 2017

Resumos de apresentações orais

Resumos de trabalhos em pôster

Resumos de trabalhos em andamento “Minuto de Loucura”

Resumos dos Minicursos, Minicursos Expressos e *Workshops* Especiais

"ESQUENTA ELC" – GRANDES TEMAS E DIÁLOGOS COM LINGUÍSTICA DE CORPUS

INTEGRATING CORPUS AND COGNITIVE LINGUISTIC APPROACHES IN LANGUAGE RESEARCH AND TEACHING: RATIONALES AND PRAXES

Palestrante: Dr. Dilin LIU (Department of English, The University of Alabama)

This speech explores the rationales and existing efforts to integrate corpus and cognitive linguistic approaches in language research and teaching. Specifically, drawing on published studies and focusing on challenging language issues, such as the use collocations, collostructions, propositions, and synonyms, the speech discusses how combining corpus and cognitive linguistic approaches may enhance the validity and reliability of research on language use and increase the effectiveness of language teaching. The discussion will particularly highlight how the combination of the two approaches may help produce new insightful understandings about language usage/acquisition issues that could not be attained otherwise. The speech ends with a brief discussion of the challenges and future directions of combining cognitive and corpus linguistic approaches in language research and teaching.

MINICURSO A - INTRODUÇÃO AO PLN/ANÁLISE DE SENTIMENTOS

Ministrante: Prof. Dr. Thiago Pardo (NILC/ICMC/USP)

Parte 1: Introdução ao PLN – Criação e anotação de *corpora*

Neste minicurso, serão apresentados os conceitos básicos da Linguística Computacional, introduzindo-se os níveis linguísticos usualmente tratados e alguns recursos, ferramentas e aplicações associados. Alguns *corpora* relevantes serão apresentados e discutidos, mostrando-se como eles têm sido usados na área.

Parte 2: Introdução à Análise de Sentimentos

Serão apresentados os conceitos básicos da área de análise de sentimentos e os desafios relacionados, que envolvem várias áreas, como Linguística, Psicologia e Computação. Alguns métodos de análise de sentimentos serão explorados, terminando com uma atividade prática sobre o assunto.

MINICURSO B – INTRODUÇÃO À LINGUÍSTICA DE CORPUS: COMEÇANDO DO BÁSICO

Ministrante: Prof. Dra. Rozane R. Rebechi (IL/UFRGS)

Na primeira parte do minicurso, apresentaremos aos participantes os fundamentos da Linguística de Corpus, partindo para um breve relato da sua aplicação para o ensino de língua estrangeira, os Estudos de Tradução e a Terminologia. Serão discutidos os conceitos de pesquisas *corpus-based* e *corpus-driven*, e apresentadas as principais ferramentas de análise de *corpora* textuais e suas funções básicas: levantamento de listas de palavras (*wordlist*), listas de palavras-chave (*keyword*) e linhas de concordância (*concord*). A segunda parte será dedicada à realização de exercícios práticos a partir de *corpora* disponíveis online para o ensino, e de *corpora* especializados de pequenas proporções para extração de terminologia e busca de equivalentes tradutórios.

MINICURSO C – COMO USAR O *SKETCH ENGINE* PARA ANALISAR O COMPORTAMENTO DAS PALAVRAS

Ministrante: Prof. Dra. Ana Frankenberg-Garcia (University of Surrey – UK)

Este *workshop* prático destina-se a todos que queiram aprender a utilizar a ferramenta *Sketch Engine* para analisar o comportamento das palavras em uma língua, variante linguística, linguagem especializada, tradução ou em qualquer texto ou conjunto de textos. Não são necessários conhecimentos prévios de *corpora*. O *Sketch Engine* é provavelmente a ferramenta de *corpora* mais avançada hoje existente. É muito simples de usar e ao mesmo tempo permite efetuar análises extremamente sofisticadas. Além de dar acesso a centenas de *corpora* em diversas línguas, prontos para serem interrogados, o *Sketch Engine* simplifica a criação de *corpora* faça-você-mesmo, ajudando a encontrar e preparar textos relevantes, a etiquetá-los com anotação gramatical e a deles extrair palavras-chave e terminologia. Após uma breve introdução, este *workshop* abordará o uso do *Sketch Engine* através de uma série de exercícios práticos. No final do *workshop*, os participantes serão capazes de: (1) utilizar as principais funcionalidades do *Sketch Engine*; (2) usar o *Sketch Engine* para analisar questões linguísticas específicas; (3) criar um *corpus* especializado; (4) extrair automaticamente palavras-chave e terminologia.

MINICURSO EXPRESSO I – A SEMÂNTICA ENCONTRA A LEXICOGRAFIA – DICIONÁRIOS BASEADOS EM *FRAMES*

Ministrantes: Grupo SemanTec (Unisinos)

Este minicurso, uma iniciativa do grupo SemanTec, propõe-se a tratar de questões relacionadas ao desenvolvimento dos dicionários Field – Dicionário de Expressões do Futebol (2014) e Dicionário Olímpico (2016). Merecerão destaque os seguintes temas: (i) Projetando um dicionário eletrônico baseado em *frames*, (ii) O uso de *corpus* eletrônico e tradução, (iii) Lexicografia Eletrônica: demandas e expectativas, (iv) Unidades lexicais: construindo a lista de verbetes e (v) A organização dos cenários e a estrutura do dicionário. Explore os dicionários do SemanTec: http://dicionariofield.com.br/ e http://dicionarioolimpico.com.br/.

MINICURSO EXPRESSO II – ESTUDOS CONTRASTIVOS EM ESPANHOL/PORTUGUÊS E LINGUÍSTICA DE CORPUS

Ministrante: Prof. Dr. Ariel Novodvorski (ILEEL/PPGEL/UFU)

O objetivo principal desta oficina é incentivar o desenvolvimento dos estudos contrastivos com base em *corpus*. Serão apresentados, de modo prático e objetivo, conhecimentos específicos para o emprego de ferramentas presentes nos programas computacionais WordSmith Tools 7.0 (Scott, 2016) e ParaConc (Barlow, 2001), além de recursos disponíveis no Corpus del Español e no Corpus do Português (Davies, 2016). Tudo será aplicado aos estudos contrastivos em espanhol/português, a partir do alinhamento e exploração de *corpora* paralelos e também do trabalho com *corpora* comparáveis, tanto do âmbito jornalístico quanto de legendas de filmes.

WORKSHOP ESPECIAL I – EXPERIÊNCIAS PRÁTICAS NA UNIVERSIDADE FEDERAL DE UBERLÂNDIA (UFU) COM LINGUÍSTICA DE CORPUS

Ministrante: Prof. Dr. Guilherme Fromm e Prof. Dr. Ariel Novodvorski (ILEEL/PPGEL/UFU)

Este *workshop* apresenta experiências de uma trajetória, construída nos últimos oito anos, na construção de um ambiente universitário favorável à multiplicação de estudos e pesquisas baseados na abordagem e nas metodologias da Linguística de Corpus. As experiências destacadas, ilustradas em atividades de iniciação científica, extensão e pós-graduação, tratam do estudo/trabalho com *corpora* tais como textos de legendas de seriados de TV em diferentes idiomas, legendas para cinema, estudos contrastivos português-inglês-espanhol e textos de literatura infantil. Trazendo o passo a passo de nossos trabalhos, buscamos inspirar atividades de aula/ensino, futuras pesquisas, monografias, TCCs, estudos-piloto e até apontar possibilidades de temas para futuros mestrados e doutorados, no cenário da formação de professores e da formação de tradutores.

WORKSHOP ESPECIAL II – AVALIAÇÃO E SIMPLIFICAÇÃO – FERRAMENTAS E PESQUISAS COM *CORPORA* PARA O ENSINO E REESCRITA DE TEXTOS

Ministrante: Prof. Ma. Aline Evers (UFRGS)

Resultados de pesquisas em Linguística de Corpus e Linguística Computacional produziram uma série de recursos capazes de qualificar o trabalho de professores e de tradutores. Quais são esses resultados? Que ferramentas computacionais disponíveis podemos utilizar atualmente para otimizar uma série de tarefas que envolvem avaliar e simplificar textos? Neste *workshop*, apresentaremos reflexões sobre as áreas de avaliação de produção textual e de simplificação de textos a partir da perspectiva da Linguística de Corpus; traremos sugestões de uso de ferramentas computacionais disponíveis com potencial de qualificar o trabalho de professores, tradutores e revisores de texto; mostraremos como ferramentas de análise textual podem nos auxiliar a fazer uma leitura diagnóstica diferenciada de textos; mostraremos como essas mesmas ferramentas também podem ser úteis em tarefas de tradução intralinguística visando à simplificação de um texto.

Workshop convidado

USING CORPORA FOR LANGUAGE LEARNING AND TEACHING

Ministrante: Dr. Dilin LIU (Department of English, The University of Alabama)

Thanks to rapid advancements in corpus linguistics and computer technology, corpora *have become an increasingly popular tool for language learning and teaching. Understanding how to use* corpora *in teaching is thus important for today's language teachers. Using information from his book on corpus use in language teaching recently published by TESOL, the facilitator introduces common useful procedures and activities for using* corpora *to teach various aspects of English, including vocabulary, grammar, and writing. Following the introduction, the participants will work both individually and in groups to develop corpus-based learning activities that promote discovery learning for their own students.*

Palestra de encerramento do ELC 2017

NEW UNDERSTANDING-INSPIRED PRACTICE IN TEACHING VOCABULARY

Palestrante: Dr. Dilin LIU (Department of English, The University of Alabama)

Vocabulary learning is very important but also challenging in L2 acquisition. Some aspects of vocabulary knowledge (such as word collocations and connotations) and some particular lexical items (such as prepositions and polysemous words) are especially difficulty due to their seemingly arbitrary usage nature or their lack of patterns. However, contemporary linguistic research has shown that there are actually clear underlying motivations and patterns in the use of these difficult English lexical aspects and items. After a quick overview of the new findings on the topic, are introduced some new research understanding-inspired activities for teaching vocabulary, including corpus-based cognitive analysis and the use of diagrams and other visuals as well as materials and activities derived from corpus data.

Resumos de apresentações orais

TOWARDS MACHINE TRANSLATION FROM BRAZILIAN PORTUGUESE-TO-LIBRAS: A CORPUS-BASED, MORPHOSYNTACTIC ANALYSIS

Francisco Aulísio dos Santos Paiva, Plínio Almeida Barbosa,
Pablo Faria, José Mario De Martino
School of Electrical and Computer Engineering – FEEC/DCA
University of Campinas (Unicamp)
Departament of Linguistics – University of Campinas (Unicamp)
aulisio.paiva@gmail.com, pabarbosa.unicampbr@gmail.com,
pablofaria@gmail.com, martino@fee.unicamp.br

Abstract: *This paper presents a methodology for identifying morphosyntactic rules for BP-Libras machine translation. Since Libras does not have a widely accepted writing system, glosses will be used as an intermediate language. The knowledge-based rules are specified on the basis of the analysis of a BP-Libras parallel corpus and from results of studies on Libras grammar applied to the sentences' pairs of the corpus.*

Resumo: Este artigo apresenta uma metodologia para a identificação de regras morfossintáticas para a tradução automática PB-Libras. Libras não tem um sistema de escrita amplamente aceito, assim glosas serão usadas como linguagem intermediária. As regras são implementadas por técnica knowledge-based a partir do exame de um *corpus* paralelo PB-Libras e estudos da gramática de Libras aplicadas aos pares de sentenças do *corpus*.

1. Introduction

Brazilian Sign Language (Libras), an official language in Brazil, is a visuospatial language of the Deaf communities. Libras is not a manual version of Brazilian Portuguese (BP), displaying its own properties as a natural language. Nevertheless, there is no widely-accepted writing system such as the Latin alphabetical system for Libras. To compensate for that, a widespread solution in the literature is the use of a gloss transcription. Glosses are words written in capital letters that represent their respective signs.

This paper presents an ongoing research targeted to identify appropriate rules for the BP-Libras Machine Translation (MT) process by using morphosyntactic information specified on the basis of the analysis of a BP-Libras parallel corpus. In this context, a gloss transcription is used as an intermediate language. The final stage of the MT process is the signing by an avatar. The BP-Libras parallel corpus contains the translation of a science textbook (De Martino et al., 2016). The translation was carried out by a team of Brazilian listeners proficient in Libras and deaf people having Libras as their first language with writing and reading skills in BP. The parallel corpus include, for each sentence of the textbook: its content in BP, a transcription in glosses, and manual and non-manual expressions. Previous known MT systems for BP to Libras are: Falibras (Brito et al., 2012), VLibras (Lima, 2015), ProDeaf and HandTalk (Vieira et al., 2014), SinaLibras 3D (Amaral, 2012). Some of the limitations of these previous approaches are: they did not systematically analyze parallel productions to derive translation rules, and some of them are limited to translate isolated words without taking into account the structure of the sentences

into which the words are inserted. Our proposal is based on the analysis of a parallel corpus BP-Libras composed of the contents of school textbooks. Currently, the parallel corpus contains the content of an elementary textbook with approximately 2,000 sentences composed of 3,000 different glosses/signs.

2. Methodology

The methodology of the present work follows the steps below.

2.1. Sentences Analysis

The pair of parallel sentences is linguistically analyzed by considering results from research in Libras grammar applied to the sentences' pairs of the corpus. This includes the SVO (Subject – Verb – Object) structures and their variations. Additionally, grammatical modalities (declarative and interrogative) are being studied in order to identify their specificities in the translation. Morphological and syntactic analyzers are used for identification of grammatical categories and syntactic constituents (Bick, 2000). For instance, the sentence "João vendeu o carro" can be glossed as "JOÃO VENDER CARRO JÁ". It is important to note that in Libras, articles or marks of gender and number are not signed, therefore they can be withdrawn in the glosses. Moreover, verbs are written in the infinitive form with tense indicated by signs referring to time.

Due to these characteristics of Libras grammar, we propose to use a lemmatizer[1] as a preprocessing phase of the translation process, enabling us to have the following analysis for the BP sentence above: João (proper noun singular) vender (third person singular, simple past tense verb) o (article) carro (simple noun singular). From that, we can get the glossed translation by removing the article and adding the gloss "JÁ" to indicate that the verb "vender" is in the past tense form (for more details on Libras transcription, see Paiva et al., 2016).

2.2. The BP-Libras translation process

We are working on the rules for translating a BP text into glosses of Libras that allows, in a later step, the control of a signing avatar. The method is, for the time being, entirely knowledge-based.

3. Results

The following examples illustrate some rules for analyzing declarative and interrogative sentences in BP and its translation to Libras' glosses.

3.1. Declarative

Input:

Carlos construiu um cata-vento.

Output:

C-A-R-L-O-S CONSTRUIR CATA-VENTO JÁ.

Algorithm Steps:

a. Parse the sentence.

[1] Maziero, E. G. (2012). Lemmatizer for Portuguese. Available: <http://conteudo.icmc.usp.br/ pessoas/taspardo/LematizadorV2a.rar>. Last access: March 3, 2017.

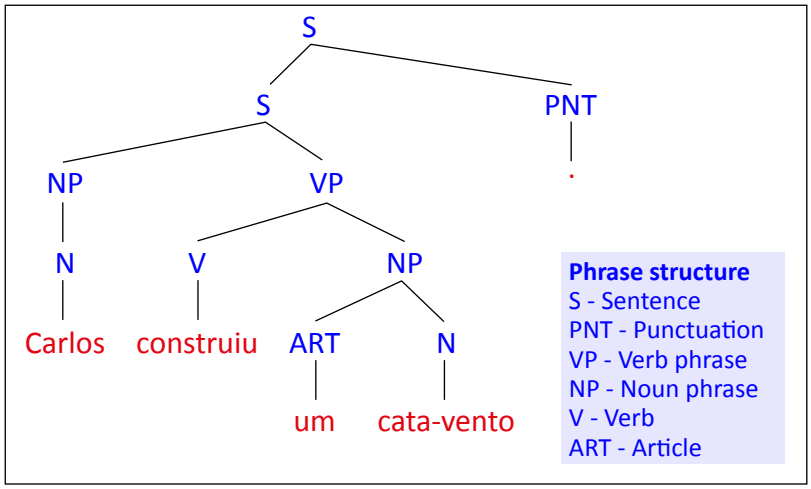

Figure 1 – Syntactic tree[2]

b. Given the above graph:

- *Recognize the modality as declarative (the punctuation sign is used to identify modality) and parse the sentence into phrases;*
- *Use a lemmatizer to retrieve lemmas, that is, the infinitive form of the verb (and its tense, person, number), articles, prepositions and other parts of speech;*
- *Delete words labeled as ART;*
- *With tense information, add "JÁ" to indicate the past;*
- *Fingerspell for proper names (Carlos case);*
- *Turn words into capital letters.*

Sometimes in Libras, the verbal tense is inferred from the context. In that case, it is not necessary to include the adverb "JÁ", as shown in the following example (Paiva et al., 2016, p. 40).

3.2. Wh-question

Input:

O que João pegou?

Output:

JOÃO PEGAR O QUE?

Algorithm Steps:

The algorithm steps are similar to the previous case, with the following changes:

[2] LX-Parser. Available: <http:// lxcenter.di.fc.ul.pt>. Last access: March 3, 2017.

a. After the syntactic analysis:

- *Recognize the modality as interrogative, which allows the use of a specific rule;*
- *Use a lemmatizer to retrieve the infinitive form of the verb, as well the grammatical categories as in the following example:*

o (article) que (pronoun) josé (proper noun singular) pegar (simple past tense verb) ?

Then, we define the following rule:

- *When article + pronoun is at the beginning of a question, move it to the end of the sentence and add a hyphen between them;*
- *With tense information, add "JÁ" to indicate the past (optional);*
- *"JOÃO" has a specific sign. It is possible to store proper nouns that have signs given by the deaf community;*
- *Turn words into capital letters.*

To better illustrate this, Figure 2 (extracted from Paiva et al., 2016) presents three snapshots of the signing by our avatar of the sentence "JOÃO PEGAR O-QUE?".

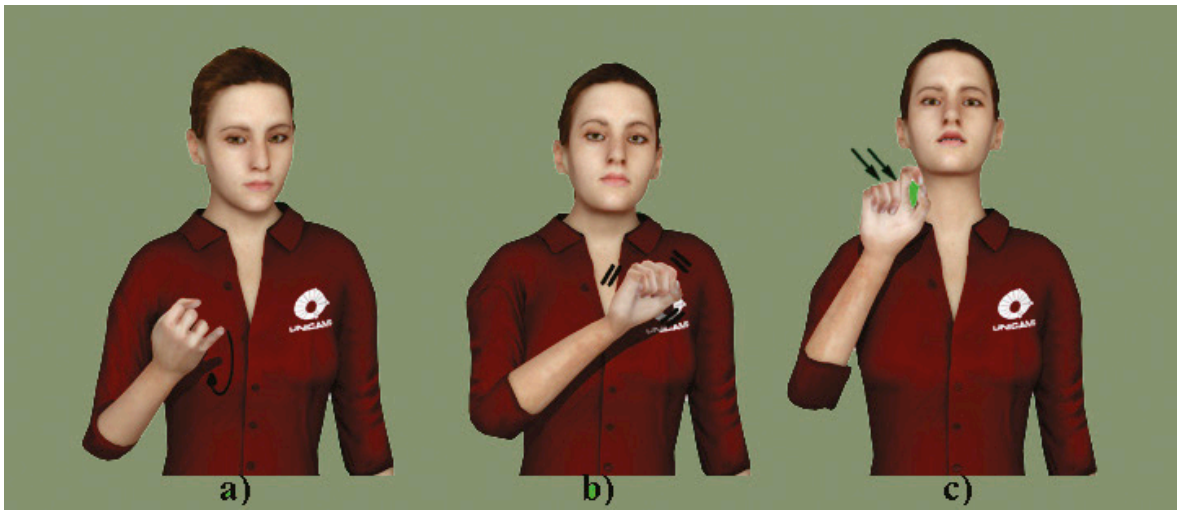

Figure 2 – Snapshots of the avatar signing a) JOÃO; b) PEGAR; c) O-QUE?

Our MT system will be able to translate sentences with similar structures. For example, the sentence of the corpus "O que a foto apresenta?" whose translation to Libras' glosses is: "FOTO APRESENTAR O-QUE?".

4. Conclusion

This work presented a methodology for implementing morphosyntactic rules that allow BP-Libras automatic translation. We studied the cases of declarative and interrogative sentences, in which the following aspects are treated: recognition of sentence modality, preprocessing of articles and verbs, rules for interrogative pronouns and verbal tense. All this favors a better knowledge of the Libras as well as stimulates social applications for deaf people.

Acknowledgments: *This work was supported by CAPES/ SDH/ MCTI No. 59/2014 – Proc. # 88887.091672 / 2014-01.*

5. References

AMARAL, W. M. ***Sistema de transição da língua brasileira de sinais voltado à produção de conteúdo sinalizado por avatares 3D***. 2012. 243 f. Tese (doutorado em Engenharia Elétrica). Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2012.

BICK, E. ***The Parsing System Palavras*** – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework , Århus, 2000.

BRITO, P. H. S.; FRANCO, N. M.; CORADINE, L. C. FALIBRAS: uma ferramenta flexível para promover acessibilidade de pessoas surdas. ***TISE: Nuevas Ideas em Informática Educativa***, 8, 2012.

DE MARTINO, J. M.; COSTA, P. D. P.; BENETTI, A. B.; ROSA, L. A.; KUMADA, K. M. O.; SILVA, I. R. Building a Brazilian Portuguese- Brazilian Sign Language Parallel Corpus using Motion Capture Data. In: WORKSHOP ON *CORPORA* AND TOOLS FOR PROCESSING *CORPORA* WORKSHOP – Collocated with PROPOR 2016 – The 12th International Conference on the Computational Processing of the Portuguese Language., 2016, Tomar. ***Proceedings Workshop on*** *Corpora* ***and Tools for Processing*** *Corpora* ***Workshop***, p. 56-63, 2016.

LIMA, M. A. C. B. ***Tradução automática com adequação sintático-semântica para Libras***. Dissertação (mestrado). Universidade Federal da Paraíba, João Pessoa, 2015.

PAIVA, F. A. S.; DE MARTINO, J. M.; BARBOSA, P. A.; BENETTI, Â.; SILVA, I. R. Um sistema de transcrição para língua de sinais brasileira: o caso de um avatar. ***Revista do GEL***, v. 13, p. 12- 48, 2016.

VIEIRA, M. C.; CORRÊA, Y.; SANTAROSA, L.; BIAZUS, M. C. V. ***Para além da interação***: o papel de aplicativos como ProDeaf e HandTalk na constituição do sujeito surdo. 3º Seminário Nacional de Inclusão Digital, Passo Fundo, 2014.

11

THE MAKING OF A BILINGUAL DICTIONARY

Ana Frankenberg-Garcia
Centre for Translation Studies – University of Surrey Guildford GU2 7HX – Reino Unido
a.frankenberg-garcia@surrey.ac.uk

Dictionaries rely on both introspection and observation to describe how words are used. Natural language processing has made it possible to observe large amounts of concentrated language data in unprecedented ways. The Oxford Portuguese Dictionary (López et al., 2015) was compiled with the aid of electronic corpora*, which have helped produce authentic, relevant entries, reflecting up-to-date usage, and natural equivalences between English and Portuguese. A summary of the compilation of the Portuguese-English side of the dictionary is presented here.*

REPRESENTAÇÕES SOCIAIS PAN-AMERICANAS: UM ESTUDO BASEADO EM TRIGRAMAS DO *GOOGLE BOOKS*

Tony Berber Sardinha, Maria Claudia Delfino, Rafael Araujo
Pontifícia Universidade Católica de São Paulo (PUCSP)
tonycorpuslg@gmail.com, claudia@fatecpg.com.br, prof.rafael_araujo@hotmail.com

Esta pesquisa teve por objetivo identificar as representações comumente construídas sobre as Américas por meio da análise dos trigramas mais recorrentes encontrados nas publicações do *Google Books* e observação da variação temática ao longo do tempo. O arcabouço teórico-metodológico desta pesquisa é a Linguística de Corpus, cuja base metodológica se fundamenta na Análise Multidimensional e, para tanto, utilizaremos os autores Biber (1988 et seq.), Biber et al. (2001; 2009) e Berber Sardinha (2014 et seq.). A base metodológica dessa pesquisa é a AMD (Biber, 1988 et seq.), definida como uma abordagem para análise de *corpus* que usa procedimentos estatísticos (principalmente análise fatorial) e que se ocupa do mapeamento das associações entre conjuntos variados de características linguísticas dentro do *corpus* de estudo (Berber Sardinha, 2004). A presente pesquisa teve o objetivo de revelar representações sociais comumente associadas aos povos do continente americano (norte-americanos, latino-americanos, central-americanos e sul-americanos) ao longo do tempo, construídas através da preferência semântica, ou seja, um agrupamento de colocações (grupo lexical) que compartilha algumas características semânticas (Hoey, 2005). Para tanto, foi realizada uma análise da coocorrência das colocações mais salientes de um *corpus* formado por listas dos trigramas mais recorrentes, extraído através da ferramenta *Google Ngrams Viewer* (Figura 1) do *Google Books* em publicações em língua inglesa entre 1800 e 2008. O *corpus* do estudo, *Pan American Corpus of Trigrams* (PACT), foi compilado e processado através de *script* desenvolvido especificamente para esta pesquisa, por meio da extração dos bigramas individuais em cada *subcorpora* (*north american, latin american, central american* e *south american*), mais seus respectivos colocados mais frequentes em livros digitalizados através da ferramenta OCR e disponibilizados pelo *Google Books*, totalizando 466.666.156.981 palavras, 2.429.890.866 páginas, e 4.520.002 publicações de 21 décadas (1800 à 2008), representando 99,6% do acervo total, uma vez que uma pequena fração dos livros foi publicada antes de 1800. Por meio da análise quantitativa pode-se interpretar os fatores em dimensões de variação (Biber, 1988 et seq.). Os resultados obtidos na análise qualitativa para confirmação das hipóteses de interpretação dos fatores, em termos de dimensões temáticas, por meio de exemplos extraídos dos textos do *corpus*, serão apresentados e ilustrados.

IDENTIFICAÇÃO DE TERMOS NO DISCURSO LITERÁRIO DE FANTASIA DA SÉRIE HARRY POTTER EM UMA ABORDAGEM DIRECIONADA POR *CORPUS*

Raphael Marco Oliveira Carneiro
Universidade Federal de Uberlândia (UFU)
raphael.olic@gmail.com

Ao propor uma perspectiva analítico-descritiva para o enfoque terminológico de unidades lexicais ficcionais, na qualidade de unidades multifuncionais ou vocábulos-termos, usadas no discurso literário de fantasia infantojuvenil da série Harry Potter, este trabalho pretende contribuir para o reconhecimento do estatuto terminológico desse tipo de unidades no escopo do desenvolvimento dos estudos terminológicos.

Em uma proposta transdisciplinar, articulamos saberes oriundos de campos diversos dos estudos linguísticos e literários na composição de nosso quadro teórico-metodológico. A transdisciplinaridade, ao contrário da lógica clássica, não está presa pela articulação de contrários (Santos, 2008), e é nessa perspectiva que concebemos as unidades lexicais. Ao aplicarmos a lógica transdisciplinar à problemática do estatuto de unidades lexicais, observamos que 'termo' é 'termo', 'termo' não é 'vocábulo', mas há um termo T, que é, ao mesmo tempo, 'termo' e 'vocábulo'. Em outras palavras, se consideramos um *continuum* de especialização de unidades lexicais, em algum ponto desse *continuum* posicionam-se unidades em funções simultâneas de termo e vocábulo, conservando características da língua comum e de linguagens especializadas. Nessa lógica, aceitam-se as articulações entre língua comum e linguagens especializadas, inclusive entre linguagem literária; a dinâmica dos movimentos entre vocábulos e termos; e a hibridização dessas duas funções em uma, vocábulos-termos, ou unidades multifuncionais.

Tendo em vista os pressupostos teóricos de quatro vertentes dos estudos terminológicos, Etnoterminologia (Barbosa, 2007), Teoria Sociocognitiva da Terminologia (Temmerman, 2000), Terminologia Cultural (Diki-Kidiri, 2009; Lara, 2007) e Terminologia Textual (Finatto, 2004; Hoffman, 2015), em articulação à semântica de mundos ficcionais (Doležel, 1998; Ryan, 2014), integramos procedimentos terminográficos direcionados por *corpus* (Barbosa, 2004; Barros, 2004; Berber Sardinha, 2004; Fromm, 2007; Krieger e Finatto, 2004), na sistematização do referencial teórico-metodológico, que nos permitiu não só compreender as especificidades dos termos ficcionais, como também gerar um glossário de possível interesse, principalmente para folcloristas.

Por termos ficcionais entendemos as unidades lexicais que, na maioria dos casos, designam elementos não pertencentes ao mundo experimentado fisicamente, à realidade fenomenológica, de forma que a existência dos elementos por elas designadas está condicionada ao texto, além de depender parcial ou totalmente da cognição; são também unidades lexicais semanticamente representativas de uma temática, usadas para a composição de um texto literário, tendo em vista a criação de um mundo ficcional. Em outras palavras, essas unidades lexicais habitam o imaginário humano e fazem parte do acervo cultural de dada sociedade.

A determinação do público-alvo do glossário partiu tanto do tipo de unidades lexi-

cais encontradas quanto do entendimento de que as manifestações do universo de discurso literário de fantasia são geralmente conhecidas como propensas a estabelecerem diálogos com discursos etnoliterários, ou seja, elas incorporam elementos tipicamente encontrados no folclore, nas lendas e nos mitos que constituem uma cultura. Isso porque as origens da fantasia podem ser traçadas a partir do folclore e da tradição oral de se contar histórias; características típicas das sociedades arcaicas que se consolidaram como base para as manifestações literárias da fantasia nas sociedades industriais e pós-industriais.

Para atingir nossos objetivos, partimos de um *corpus* de estudo composto pelos sete volumes da série literária Harry Potter e de outros três volumes (*Fantastic Beasts and Where to Find Them, Quidditch Through the Ages, The Tales of Beedle the Bard*), em inglês, que detalham o mesmo mundo ficcional criado por J. K. Rowling. Esse *corpus*, quando processado pelo programa WordSmith Tools 6.0 (Scott, 2012) e suas três ferramentas, Concord, KeyWords e WordList, permitiu a identificação dos termos e o acesso aos seus contextos linguísticos de ocorrências.

Ressaltamos que o conceito de *corpus* nesta pesquisa é de particular relevância, dada à abordagem direcionada pelo *corpus*. Em vez de partir de pressupostos anteriores ao desenvolvimento da pesquisa em relação ao tipo de unidades lexicais que comumente recebem o estatuto de termos, buscamos deixar que as próprias evidências do *corpus* falassem por si mesmas. O benefício dessa abordagem foi verificado quando da análise das palavras-chave e de seus contextos linguísticos. Isso porque pudemos identificar unidades lexicais típicas de um universo semântico caracterizado por uma temática do universo literário de fantasia infantojuvenil, ou seja, Magia e Bruxaria (*Witchcraft and Wizardry*). A partir dessa identificação, extraímos os contextos linguísticos dessas unidades nos quais encontramos elementos definitórios característicos de definições formais simples, semiformais e complexas (Pearson, 1998), o que em conjunto com o recorte semântico-conceptual destacado nos possibilitou atribuir a essas unidades lexicais o valor de termos dentro do universo de discurso literário de fantasia.

Como parte da metodologia, também descrevemos elementos da macroestrutura e da microestrutura textuais, construímos uma representação da organização conceptual subjacente à temática das obras e elaboramos uma ficha terminológica, preenchendo quinze fichas para demonstrar a viabilidade de nossa proposta. Essas fichas integram informações referentes a quinze termos: *Avada Kedavra, Deathly Hallows, Dementors, expecto patronum, Grim, Horcrux, house-elf, Muggle, Muggle-born, owl, phoenix, Sectumsempra, Thestrals, wand, wizard.*

Concluímos que os termos usados no discurso literário de fantasia infantojuvenil, conforme manifestado nas obras de Harry Potter, apresentam especificidades no interior do universo de discurso literário de fantasia e interdiscursividade com o universo de discurso etnoliterário, como o folclore. As unidades lexicais ficcionais atualizam estatuto terminológico no discurso literário de fantasia infantojuvenil da série Harry Potter devido aos seguintes aspectos: elas fazem parte de um sistema conceptual estruturado dentro de uma temática específica, *Witchcraft and Wizardry*; atuam na composição de um mundo ficcional semioticamente construído pela força modelizante da linguagem literária; possuem intertextualidade e interdiscursividade intra e interuniverso de discurso com discursos etnoliterários; atualizam um sistema de valores em investimentos axiológicos positivos e negativos; designam conceitos formados com semas do universo de

discurso em que são usadas; referem-se aos particulares de um mundo ficcional; quanto à função simbólica, atuam no plano do imaginário, de maneira que é nas narrativas ficcionais que encontramos as razões para conceber as relações simbólicas.

Acreditamos que estudos como o desenvolvido possam ampliar o escopo dos estudos terminológicos e fazê-los mais responsivos a diferentes manifestações textuais passíveis de fazerem uso de unidades lexicais com estatuto terminológico. Nesse sentido, a abordagem direcionada por *corpus* é fundamental para a exploração de manifestações textuais que possam revelar fenômenos terminológicos previamente não concebidos.

38

CLUSTERING AND HIERARCHICAL ORGANIZATION OF OPINION ASPECTS: A CORPUS STUDY

Francielle Alves Vargas e Thiago Alexandre Salgueiro Pardo
University of São Paulo (USP)
francielleavargas@usp.br; taspardo@icmc.usp.br

This paper consists of an empirical study on the problem of clustering and hierarchically organizing opinion aspects in product reviews in order to support aspect-based opinion mining applications. We performed a corpus study for characterizing and understanding the involved tasks, looking for linguistic patterns and convergences and divergences across domains. The process has been manually performed and resulted in reference data for future developments and evaluation of automatic methods in the area.

43

REVISTA BRASILEIRA DE LINGUÍSTICA APLICADA: MULTIDIMENSÕES TEMÁTICAS

Maria Antonieta Celani, Tony Berber Sardinha, Maria Claudia Delfino, Marianne Rampaso, Rafael Araujo e Ulysses Diegues
Faculdade de Tecnologia de Praia Grande
claudia@fatecpg.com.br

Com o intuito de revelar os principais temas mais recorrentes e de maior interesse nas publicações da Revista Brasileira de Linguística Aplicada (RBLA) ao longo do tempo – essa pesquisa tem como objetivo identificar as dimensões temáticas – espaços linguísticos – que refletem os assuntos de interesse e tratados com maior frequência na Linguística Aplicada (LA) por meio da investigação do léxico dos textos, além de verificar se existe uma relação entre estes temas e o período de tempo desde a primeira publicação da revista. Essa pesquisa se fundamenta por ser a publicação em questão um dos principais veículos de disseminação do conhecimento construído na LA brasileira e uma vez que, até a presente data, não existem muitos estudos que se debrucem sobre a variação lexical e temática desta área do conhecimento, a presente pesquisa se justifica e pretender preencher esta lacuna.

45

ADDING INFORMATION IN ARGUMENTATIVE TEXTS: A LEARNER CORPUS-BASED STUDY OF ADDITIVE LINKING ADVERBIALS

Jessica Maria da Silva Queiroz, Jessica Ceritello Alves e Deise Prina Dutra
Universidade Federal de Minas Gerais (UFMG)
jessicaqueiroz23@hotmail.com, jessicaceritello@gmail.com, deisepdutra@gmail.com

The possibilities for studying human languages offered by corpus linguistics (CL) are endless. Since its emergence as a promising methodology, CL has developed into a "heterogeneous field" (McEnery; Hardie, 2012, p. 1) that allows linguists to collect and explore in different levels a large amount of authentic language data. The kind of data compiled depends on the purpose of the research, but it is usually written or spoken and representative of native or learner language. In our study, we deal with learner and native language, but focus primarily on learner corpora*, which can be defined as "electronic collections of foreign or second language learner texts assembled according to explicit design criteria" (Granger, 2009, p. 14). This type of corpus is an attempt to apprehend the language produced by learners from one or various proficiency levels. Similarly to any native corpusbased study, investigations based on learner* corpora *have unraveled several features regarding the lexis, syntax, morphology, phonology, discourse, and phraseology of learner language (see Coxhead, 2010 for an overview). Aiming to develop a contrastive interlanguage analysis (CIA) (Granger, 1996; 1998), we have selected the argumentative essays written by B1 (intermediate) level university students from two Brazilian learner* corpora*: the Corpus of English without Borders (CorIsF), and the Corpus of English for Academic Purposes (CorIFA). The two subcorpora contain 48,442 and 81,586 word tokens, respectively. In addition, a native English corpus was used for comparative purposes: the Louvain Corpus of Native English Essays (LOCNESS). LOCNESS was chosen as it contains a significant and comparative subcorpus of argumentative essays with 228,920 word tokens, written by American and British university students.*

55

REPORTING ON THE DEVELOPMENT OF SKETCH GRAMMAR FOR ACADEMIC PORTUGUESE

Tanara Zingano Kuhn, Iztok Kosem
CELGA-ILTEC, Universidade de Coimbra, Portugal e Faculdade de Letras da Universidade de Lisboa, Portugal. Capes scholarship holder – Brasil
Trojina, Institute for Applied Slovene Studies, Liubliana, Eslovênia & Faculty of Arts, University of Ljubljana, Eslovênia
tanarazingano@outlook.com, iztok.kosem@trojina.si

Automation of dictionary-making has recently reached the point envisaged by Rundell and Kilgarriff (2011) in which data are automatically extracted from the corpus and imported into dictionary writing system, where then lexicographers analyse and edit the information. This was the method adopted in the conceptualisation and compilation

of a model for an online corpus-driven dictionary of Portuguese for university students. At the heart of the procedure of automatic extraction of data from the corpus are Word Sketches – "a one-page, corpus-based summary of a word's grammatical and collocational behaviour" (Kilgarriff et al., 2004, p. 105), which are a feature of the Sketch Engine *tool (Kilgarriff et al., 2004). In order to build word sketches, two conditions have to be met. One is a POStagged corpus, and the other is sketch grammar. This contribution reports on the development of a new sketch grammar for academic Portuguese. First, we provide a brief presentation of the corpus. Next, we offer a succinct description of the sketch grammar development, drawing attention to some of the main problems encountered and workarounds proposed. Finally, we summarise the main findings, highlight important implications of our research, and point out suggestions for further improvement of the sketch grammar.*

Given that existing corpora *and sketch grammars for Portuguese were unsuitable for our purposes, we had to compile a corpus of academic Portuguese, and, based on it, devise a completely new sketch grammar.*

The newly compiled corpus CoPEP – Corpus de Português Escrito em Periódicos *(Corpus of Written Portuguese in Academic Journals) (Kuhn e Ferreira, 2016) contains around 10,000 texts totalling over 40 million words extracted from academic journals published on the Brazilian and Portuguese national collections of SciELO (Scientific Electronic Library Online). The texts are distributed among three Schools of Knowledge, and further divided into six Great Areas. CoPEP is a synchronic corpus, the vast majority of its texts having been published between 2000 and 2016 (only 2% of texts are from the 1990s). The subcorpora for each language variety are of almost the same size and consist of a similar number of words per both Great Areas and Schools, making the corpus evenly balanced.*

Sketch grammar is a file with grammatical relations, or gramrels, and processing directives for the Sketch Engine *system to compute different types of relations through statistics calculations. The data obtained with these computations then form the basis of the Word Sketch feature in the* Sketch Engine*, and relatedly, the Thesaurus and Sketch Diff features. Sketch grammars devised for POS-tagged* corpora *use regular expressions over POS-tags to find matches for grammatical relations. Queries are written in Corpus Query Language (CQL), with attribute-values names following the tagset used for corpus tagging.*

Devising the sketch grammar for academic Portuguese (henceforth AcadPortSkG) consisted of two phases: 1. writing gramrel queries utilising a smaller sample corpus and 2. evaluation of the performance of the sketch grammar devised in the first phase on the entire CoPEP corpus.

The first phase adopted a trial and error method, where queries were written and tested several times until satisfactory results were reached. Then, we moved to the evaluation of the resulting sketch grammar on the entire CoPEP data with the objective of verifying whether the devised gramrel queries captured correct information. This second phase thus entailed compiling the CoPEP corpus in the Sketch Engine *using AcadPortSkG, defining a methodology of evaluation, conducting the evaluation, and proposing workarounds for gramrels in which annotation problems seriously affected the results.*

The current version of AcadPortSkG comprises symmetric (1), unary (5), dual (14), and trinary (2) grammatical relations covering attributive (pre- and postpositional) and predicative adjectives; nouns as predicative complement, subjects, and objects of verbs

(unmarked order); prepositional phrases with nouns and verbs; infinitive as verb/noun complement; impersonal and personal verbal passive constructions; verbal constructions with pronoun se*; verbs followed by clauses introduced by* que *and other conjunctions; verbs with gerund as a complement; and adverb-verb and adverb-adjective pairs.*

This new sketch grammar for academic Portuguese has had implications not only for our work on the dictionary of Portuguese for university students, but also for Portuguese corpora *in general. A comparison with the default sketch grammar available for Freeling-tagged* corpora *of Portuguese revealed that AcadPortSkG comprises a larger number of grammatical relations for nouns, verbs and adjectives, and many new rules for adverbs, thus broadening up word class coverage. In addition, the queries of existing sketch grammars, which were consulted in developing AcadPortSkG, were adapted and now yield better results. Lastly, AcadPortSkG contains queries which were carefully devised in a way to overcome detected annotation errors, making the sketch grammar more accurate.*

Especially important is that this new grammar yields very rich results, both in terms of quality and quantity – most of the times, it produces more data than can be handled by a human, which is in fact typical of sketch grammars for automatic extraction of lexical data (Kosem et al., 2013). The proposed dictionary of academic Portuguese can now take maximum advantage of this new approach in lexicography, meaning that lexicographers can get vast amounts of structured information directly into the dictionary-writing system. The richness of the lexicographic evidence obtained from the corpus due to the underlying sketch grammar is thus manageable, enabling compilation of more accurate dictionary entries.

AcadPortSkG can also be used with any other corpus of Portuguese tagged with Freeling 3.0. Such corpora *will benefit from our sketch grammar on two levels. In terms of their exploration, users of the* corpora *will be able to conduct a more thorough and reliable lexical analysis due to a greater number of grammatical relations and their (improved) accuracy. Concerning the development of tools for Portuguese, the very process of sketch grammar development has revealed problems with corpus annotation, and this can be used in finding solutions to improve the Freeling tagger and informing other resource developers of potentially problematic areas.*

We will continue to improve the sketch grammar, but as the current version already yields very good results, for both academic and general Portuguese, we decided to make it available in the Sketch Engine *for researchers using/making Freeling-tagged* corpora *of Portuguese.*

65

POMMERSCHE KORPORA: UM CONJUNTO DE *CORPORA* DIALETAIS DA VARIEDADE BRASILEIRA DO POMERANO

Neubiana Silva Veloso Beilke
Universidade Federal de Uberlândia (UFU)
neubeilke@hotmail.com

Introdução: A nossa pesquisa de mestrado teve como objetivo principal realizar a compilação de *corpora* do pomerano no Brasil. O pomerano é uma variedade linguística

germânica, pois pertence ao tronco indo-europeu e à família das línguas germânicas, estando situado dentro do grupo do baixo-alemão, proveniente das terras planas do norte da Europa. Referimo-nos especificamente ao Pommersches Plattdeutsch (Vollmer, 2008; Herrmann-Winter, 1998, 2003, 2013), ou seja, ao baixo-alemão pomerano. Em nosso trabalho, optamos por adotar variados tipos de fontes para a composição de um conjunto formado por *corpora* escritos (constituído por diários, músicas, poemas, receitas, livros, etc.) e por um *corpus* oral, composto a partir de entrevistas sociogeolinguísticas que foram realizadas com falantes de pomerano e, posteriormente, transcritas, a fim de contribuir para a documentação do pomerano.

Objetivos: Nosso objetivo geral foi compilar materiais autênticos provenientes do dialeto pomerano no Brasil para constituição de um banco de dados linguísticos (*corpus*). Nossos dois objetivos específicos foram identificar a sobrevivência ou o desaparecimento do pomerano nas regiões alvo de nossa pesquisa – no entorno do Vale do Rio Doce/MG e no interior do Rio Grande do Sul – e comparar brevemente o pomerano coletado nessas duas regiões.

Fundamentação Teórica: Para alcançar os referidos objetivos, o estudo teve como fundamentação teórica a Lexicologia e a Linguística de Corpus (LC), precisamente, as teorias que fundamentam a abordagem-metodologia da LC. Ademais, dada a natureza interdisciplinar desta pesquisa, contemplamos também as noções da Sociogeolinguística (SGL), campo de estudo que subsidiou nossa coleta de dados orais. Os princípios do empirismo e da perspectiva descritiva também inspiraram nosso trabalho. Entre nossos referenciais, encontram-se autores como Biderman (1992, 2001); Zavaglia e Welker (2013); Sinclair (1991, 2004); Fillmore (1992); Perini (2006, 2008); Berber Sardinha (2004); Tognini-Bonelli (2001); Parodi (2010); Finatto e Novodvorski (2014); Raso e Mello (2009); Mello (2012); Araújo (2012); Cristianini (2007, 2012); Takano (2013); Von Borstel (2014), dentre outros. Após pesquisas e reflexões bibliográficas, definimos que, para nós, a LC é uma abordagem-metodologia de princípios descritivos, que se fundamenta em dados autênticos e se relaciona com as evidências de maneira ampla, utiliza metodologia de pesquisa empírica, permite a produção de conhecimentos variados fundamentados na realidade linguística, além de nos guiar na investigação de hipóteses não premeditadas e na descoberta e comprovação de fatos linguísticos.

Hipóteses: Formulamos quatro hipóteses para esta pesquisa: i) a cogitação de que o pomerano não é ágrafo, pois a coleta e compilação de *corpora* demonstra a existência de formas escritas; ii) a ideia de que os falantes são encontrados, em grande maioria, no meio rural, podendo ser inexistente, em algumas localidades, a presença de falantes de pomerano na zona urbana; iii) a verificação de interferências de caráter linguístico sofridas pelos pomeranos em contato com outras etnias; e iv) a consideração de que o contato dos descendentes de pomeranos com o português tenha permitido o surgimento de uma variedade brasileira do pomerano. Dentre as quatro hipóteses, três foram confirmadas, tendo sido negada apenas aquela que se refere à presença de falantes de pomerano limitados às zonas rurais, pois foi possível encontrar falantes em zonas urbanas, como Itueta/MG e Canguçu/RS.

Metodologia: Nossa metodologia, de modo geral, envolveu a coleta, compilação, transcrição e tratamento de dados escritos e orais, bem como a convenção e a conversão da escrita em um padrão único (uniformizado, dentro do contexto da nossa pesquisa) para as amostras linguísticas pomeranas. Para a coleta de dados orais, adaptamos o Questionário Sociolinguístico (QS) e o Ques-

tionário Semântico Lexical (QSL), este último para a versão em alemão Lexikalisch-Semantischer Fragebogen (LSF); selecionamos algumas localidades para realizarmos visitas prévias e pesquisas a fim de fazer a caracterização histórico-geográfica das localidades, conforme a Sociogeolinguística (SGL); organizamos um plano de recrutamento com critérios de inclusão e exclusão dos candidatos às entrevistas, segundo as variáveis da SGL e as exigências do CEP (Comitê de Ética em Pesquisas com Seres Humanos); realizamos, também, procedimentos para a efetivação das entrevistas, como, por exemplo, a preparação dos instrumentos de coleta e testagem do método por meio da realização de entrevistas piloto. Fizemos a coleta nas seguintes localidades: Vila Neitzel/MG; Itueta/MG; Arroio do Tigre/RS; São Lourenço do Sul/RS; Canguçu/RS e Santa Maria de Jetibá/ES. Decidimos realizar coleta de dados em Santa Maria de Jetibá/ES, por ser conhecida como a capital pomerana do Brasil e também para permitir futuras comparações com o pomerano encontrado em outras localidades. Após a coleta, efetuamos o tratamento dos dados orais, quando foi necessário desenvolver um método de transcrição, o qual foi feito com base na ortografia germânica do dicionário Duden e da bíblia pomerana (a Barther Bibel), e também realizamos os procedimentos de etiquetagem parcial e de exclusão de dados antroponímicos. Ainda sobre os procedimentos metodológicos para a constituição dos *corpora* pomeranos, organizamos todo o material compilado como uma unidade, processo durante o qual realizamos a codificação, nomenclatura, agrupamento, separação e reagrupamento dos dados, segundo suas classificações em diamesias ou gêneros ou domínio discursivo ou suportes. Neste trabalho, utilizamos o OmniPage, o ArcGis, o WordSmith Tools e o SubExtractor como recursos tecnológicos. Também elaboramos uma listagem dos textos que constam nos *corpora*, bem como fizemos arquivamento de segurança do conjunto dos *corpora* via *drive* virtual online.

Resultados: Ao final, obtivemos um conjunto de dados nomeado *Pommersche Korpora* (PK), que contém quatorze *corpora* escritos, a saber: o (i) *Corpus* de Inscrições dos Túmulos Pomeranos; o (ii) *Corpus* de Livros de Registros Eclesiásticos; o (iii) *Corpus* de Cartas Pessoais; o (iv) *Corpus* de Receitas; o (v) *Corpus* do Jornal Folha Pomerana; o (vi) *Corpus* de Textos Diversos da Internet; o (vii) *Corpus* de Legendas de Documentários; o (viii) *Corpus* de Trabalhos Acadêmicos; o (ix) *Corpus* de Textos Religiosos; o (x) *Corpus* de Músicas Pomeranas; o (xi) *Corpus* Literário Pomerano; o (xii) *Corpus* de Livros; o (xiii) *Corpus* de *Sprüche* Diversos; o (xiv) *Corpus* de Palavras Soltas; e um *corpus* oral, a saber, o (xv) *Corpus* Oral de Entrevistas Interativas. O *Pommersche Korpora* Escritos (PKE) totalizou 79.290 *tokens* e 15.515 *types*. O *Pommersch Korpus* Oral (PKO) totalizou 50.376 *tokens* e 7.309 *types*. O conjunto de todos os *corpora* – o *Pommersche Korpora* ou PK – em todas as suas diamesias, modalidades e gêneros – constitui um acervo estatístico de 129.666 *tokens* e 20.672 *types*, o qual foi classificado como de tamanho pequeno-médio (conforme parâmetro de Berber Sardinha, 2004) e contém dados que permitem identificá-lo como um conjunto de *corpora* dialetais multilíngues contatuais, visto que reúne léxico pomerano autêntico, tanto do dialeto ativo quanto de textos históricos, contendo inclusive amostras linguísticas em pomerano, em alemão-padrão, em hunsriqueano e em português, todos dentro de um mesmo texto. Avaliamos que, por meio dos resultados alcançados em nosso trabalho de mestrado, conseguimos preencher a lacuna no que diz respeito à ausência de pesquisas acerca de regiões (como, por exemplo, a região leste de Minas Gerais) com presença de fala pomerana no Brasil e também à ausência de crítica ao trabalho de Tressmann (2008); apresen-

tamos documentos históricos para abordar a imigração pomerana; debatemos a questão acerca do pomerano ser língua ou dialeto; desenvolvemos procedimentos de tratamentos de dados orais dialetais e um método de consulta à bíblia pomerana de 1588; ampliamos o leque de fontes de coleta de dados linguísticos e fragmentos dialetais; fizemos transcrições de documentos históricos e fontes primárias (inclusive em letra gótica). Quanto à análise qualitativa dos resultados alcançados, esboçamos análises prévias sobre fatos linguísticos direcionados pelos *corpora*. Verificamos, com base nas evidências encontradas no PK, a existência de um pomerano que apresenta influências portuguesas, alemãs e dialetais (de outras variedades germânicas), por isso o chamamos de *Brasilianisch-Pommersch*, ou seja, uma variedade brasileira do pomerano. Encerramos nosso trabalho lançando perspectivas futuras para o estudo do pomerano, com sugestões diretas de desdobramentos futuros possíveis por meio do trabalho iniciado, tais como: expansão dos *corpora* do PK; transformar o PK em monitor; etiquetagem completa; estudo dos verbos e substantivos mais frequentes nos *corpora*; estudo dos dados do ponto de vista do contato de línguas; produção de um atlas linguístico parcial; produção de materiais didáticos com base em evidências empiricamente coletadas (evitando a formulação de frases artificias para exemplificações); utilização dos dados como ferramenta auxiliar na tradução pomerano-alemão-padrão; levantamento de hipóteses sobre a sintaxe pomerana; estudos de perdas de alguns fonemas e grafemas nas posições finais das palavras; auxílio na produção de obras lexicográficas; análise sociolinguística com base nas respostas obtidas por meio da aplicação do QS; comparação com os dados de outros *corpora* (por exemplo, *Corpus of Historical Low German*/*Corpus* do baixo-alemão histórico e/ou *Referenzkorpus Mittelniederdeutsch*/*Corpus* de referência do baixo-alemão médio). Portanto, acreditamos e almejamos ter deixado uma contribuição para a compilação de outros *corpora* dialetais.

67

A ESCRITA DE PAULO COELHO: DO VOCABULÁRIO À COMPLEXIDADE TEXTUAL

Vinícius Alces Machado
Universidade Federal do Rio Grande do Sul (UFRGS)
viniciusalces@gmail.com

São descritas as escolhas lexicais de um conjunto de obras do escritor e místico Paulo Coelho (PC), tomado como *corpus* de estudo, em contraste com obras de Augusto Cury (AC), autor conhecido no estilo de literatura de autoajuda. Como *corpus* de referência, utiliza-se o *Corpus* Brasileiro (Berber Sardinha, 2013), complementado com a lista de entradas do Dicionário Ilustrado de Português de M.T.C. Biderman, um dicionário associado ao universo vocabular de estudantes do Ensino Fundamental. Esse dicionário serve como uma referência para verificação de escolhas lexicais compatíveis, em tese, com público de menor escolaridade e pouco hábito de leitura. Os dados dos *corpora* de estudo e contraste foram gerados com os programas AntConc (Anthony, L., 2014), Flesch Calculator (Wilkens, 2016) e Compare two lists (Whitehead BaRC Public Tools, 2013). Os resultados mostram que a maioria das mil palavras mais frequentes das obras

de PC integram o dicionário de Biderman, o que, somando-se ao fato de a palavra "QUE" ocupar a primeira posição das palavras mais frequentes, permite supor uma escrita próxima da oralidade e vocabulário acessível para leitores pouco experientes. Semelhante ocorre com AC, sendo "de" a palavra mais frequente, o que permite refletir que o uso de um vocabulário considerado menos complexo seja importante no apreço do público por essas obras.

68

COLOCAÇÕES ESPECIALIZADAS NA ÁREA DO DIREITO COMERCIAL INTERNACIONAL E PROPOSTA DE GLOSSÁRIO TRILÍNGUE

Jean Michel Pimentel Rocha e Adriane Orenha-Ottaiano
Universidade Estadual Paulista (UNESP); Universidade Estadual Paulista (UNESP)
jeanpimentel_sp@yahoo.com.br, adriane@ibilce.unesp.br

Esta pesquisa, ancorada no aporte teórico da Linguística de Corpus e da Fraseologia, teve por objetivo desenvolver um estudo teórico-metodológico acerca dos procedimentos necessários ao levantamento e à análise sintático-morfológica, léxico-semântica e tradutológica das colocações especializadas extraídas do *corpus* paralelo, em inglês e em espanhol, constituído pelos anuários (1968-2010) da UNCITRAL (Comissão das Nações Unidas para o Direito do Comércio Internacional); e de dois *corpora* comparáveis em português: um compilado pela ferramenta BootCat Front End (Zanchetta, Baroni e Bernardini, 2011) e, outro, pela coleta de textos online da área do Direito Comercial Internacional. Com base neste estudo, determinamos as colocações funcionalmente equivalentes (Tognini-Bonelli e Manca, 2004) em português e elaboramos uma proposta de glossário trilíngue nas direções tradutórias inglês-espanhol-português, na área supracitada.

70

FRAMES DE COMPREENSÃO E *CORPORA*: ESTUDO DE CASO COM USO DA FERRAMENTA *SKETCH ENGINE*

Aline Nardes dos Santos e Rove Chishman
Universidade do Vale do Rio dos Sinos (Unisinos)
aline.nardes@gmail.com, rove@unisinos.br

Este trabalho aborda o uso da ferramenta *Sketch Engine* (Kilgariff et al., 2004) na descrição linguístico-cognitiva de *frames* de compreensão (Ziem, 2014), caracterizados como estruturas de conhecimento esquemáticas ativadas pelos falantes conforme seus conhecimentos de mundo. O estudo de caso apresentado mostra como o uso do *Sketch Engine*, principalmente da ferramenta Word Sketch, facilita a identificação de evocadores potenciais do *frame* analisado.

O QUE AS RESENHAS DE VIAJANTES PODEM NOS DIZER SOBRE A CULTURA? A LINGUÍSTICA DE CORPUS AJUDA A RESPONDER

Sandra Navarro
Universidade de São Paulo (USP)
sandranavarro@usp.br

Este trabalho tem como objetivo investigar as diferenças culturais entre o Brasil e os Estados Unidos, por meio da análise linguística baseada em um *corpus* de resenhas publicadas por viajantes brasileiros e americanos no site TripAdvisor, interpretando os resultados a partir das teorias sobre orientações culturais (Hall, 1977; Walker et al., 2003; Katan, 2004). O objetivo deste estudo está em acessar as orientações culturais por meio da observação dos padrões da linguagem. Adotamos, portanto, a perspectiva postulada pela Linguística de Corpus, segundo a qual a linguagem é constituída de padrões não aleatórios e consagrados pelo uso, que recebem significado a partir de determinado contexto (Sinclair, 1991). O *corpus* construído para este estudo contém resenhas de viajantes retiradas do *site* TripAdvisor e escritas em inglês por americanos e em português por brasileiros. Esse conteúdo está subdividido em quatro categorias: resenhas de americanos sobre hotéis nos EUA e hotéis no Brasil e resenhas de brasileiros sobre hotéis nos EUA e hotéis no Brasil. O *corpus* está balanceado, com 5.000 resenhas em cada idioma, 892.085 palavras em inglês e 499.094 palavras em português, totalizando 10.000 resenhas e 1.391.179 palavras. Os achados descritos apontam para diferenças culturais em termos de orientação de pensamento (Walker et al., 2003). A cultura americana privilegia o pensamento linear e indutivo. Essa forma de raciocínio enfatiza fatos concretos, estatísticas, medições, exatidão, privilegia a lógica causa e efeito. Assim, ao descrever uma experiência, o americano busca formas concretas de expressar sua avaliação, por exemplo, estabelecendo um ponto de referência, um "padrão", o que explica o uso recorrente de 'standards'. Por sua vez, a cultura brasileira tende ao pensamento sistêmico e dedutivo, que enfatiza uma abordagem integrada ou holística. Esse ponto de vista concentra-se menos nos fatos concretos e mais nas conexões entre as partes, com o uso frequente de analogias, metáforas, símiles. Isso explica a ausência de referências recorrentes a um "padrão" determinado como critério para avaliações. Os dados também podem ser interpretados em termos de orientação de ação (Walker et al., 2003). Essa orientação distingue entre duas formas de motivação, ser e fazer. As culturas voltadas ao fazer, como a americana, não encontram dificuldade em separar fatos de sentimentos pessoais. Assim, é possível criticar a ação sem que isso implique necessariamente uma crítica à identidade da pessoa. Em contraste, em culturas orientadas ao ser, como a brasileira, uma crítica é facilmente compreendida como um ataque à identidade da pessoa. Essa diferença pode ser confirmada pelos distintos usos de 'Brazilian' e 'brasileiro' no *corpus*. Ao avaliar como inferior a qualidade de alguns hotéis brasileiros, os americanos pautaram suas opiniões em padrões objetivos, daí a colocação 'by Brazilian standards', o que indica uma realidade de mercado e não necessariamente um ataque à nacionalidade brasileira. Para o brasileiro, é difícil fazer essa distinção. Constatou-se que uma pesquisa que visa comparar padrões culturais a partir de padrões da linguagem pode produzir resultados que enriquecem nossa consciência cultural, não apenas em relação a uma cultura estrangeira, mas em relação à nossa própria.

ELABORAÇÃO DE UM *CORPUS* DE COMUNICAÇÕES PILOTO-CONTROLADOR EM LÍNGUA INGLESA – QUESTÕES TEÓRICAS E METODOLÓGICAS

Patricia Lucks
Universidade Estadual Paulista Júlio de Mesquita Filho
patricialucks@gmail.com

A aviação civil tem se desenvolvido e aumentado exponencialmente nas últimas décadas, tanto no Brasil quanto em todo o mundo. Porém, apesar de se tratar de uma área que exige os mais rígidos padrões de treinamento de pessoal e de manutenção de equipamentos, esse crescimento em larga escala também tem feito aumentar o número de incidentes e acidentes aéreos, que podem ser causados por uma combinação de diferentes tipos de falhas humanas e/ou tecnológicas. Dentre elas, existe a possibilidade de falha na comunicação entre pilotos e controladores de tráfego aéreo, o que pode ocorrer em virtude de deficiências na proficiência linguística desses profissionais. Há diversas pesquisas que apontam problemas nas comunicações como fatores contribuintes de acidentes aéreos (cf. Cushing, 1997; Prinzo e Morrow, 1999; Eurocontrol, 2006; Monteiro, 2009). Assim, a proficiência em língua inglesa é considerada uma ferramenta fundamental para a segurança das comunicações aeronáuticas que envolvem o tráfego aéreo internacional, sobretudo porque os profissionais precisam utilizar o inglês como língua franca para se comunicar com falantes nativos e não nativos, das mais variadas nacionalidades. Para garantir essa proficiência, a Organização de Aviação Civil Internacional (OACI) publicou, em 2004, um documento conhecido como Doc 9835 (OACI, 2004; 2010) que estabelece os requisitos para a capacitação e a avaliação de pilotos e controladores que atuam em tráfego aéreo internacional, a fim de que esses profissionais tenham competência para se comunicar em língua inglesa com segurança, principalmente em situações em que seja necessário extrapolar o uso da Fraseologia Aeronáutica padrão e utilizar a língua inglesa comum (*plain English*). Contudo, por se tratar de algo muito recente, são ainda escassos os estudos acadêmicos sobre essa linguagem tão peculiar, bem como sobre a eficiência dos requisitos da OACI para de fato assegurar a segurança nos céus. Dentre as pesquisas já realizadas, destacamos alguns estudos que se apoiaram na Linguística de Corpus (cf. Bocorny, 2008; Sarmento, 2008; Alizieri, 2010; Lopez, 2013; Prado, 2015). Esses trabalhos são fundamentais para uma melhor descrição e compreensão da linguagem aeronáutica, pois demonstram que é necessário que a OACI reveja algumas determinações de seus documentos normativos (OACI, 2004, 2009, 2010). A fim de oferecer uma contribuição para esses estudos, neste trabalho, apresentamos a compilação de um *corpus* eletrônico oral a partir da coleta e transcrição de gravações das comunicações rotineiras e não rotineiras, em língua inglesa, entre pilotos e controladores de tráfego aéreo do Sistema de Controle do Espaço Aéreo Brasileiro (SISCEAB). O principal diferencial em relação ao *corpus* aqui apresentado (chamado de *corpus* SISCEAB) e os outros já existentes é que este se trata do primeiro *corpus* compilado no Brasil nesses moldes, 100% composto por interações realizadas no espaço aéreo brasileiro, o que possibilitará estudos sobre controladores brasileiros que são falantes de inglês, com diferentes níveis de proficiência, e suas facilidades e dificuldades de compreensão diante de falantes de inglês de diversas nacionalidades,

também com diferentes níveis de proficiência. As gravações foram coletadas durante dois anos, em localidades com fluxo intenso de aeronaves comerciais civis internacionais, a fim de compor uma amostra suficientemente representativa das situações alvo em diferentes regiões do Brasil, nos três órgãos operacionais que se comunicam com aeronaves durante as diferentes fases do vôo (Torre de Controle, Controle de Aproximação e Centro de Controle de Área). O método para transcrições foi o sugerido por Raso e Mello (2012) para linguagem oral e, para questões específicas da aviação, analisamos as soluções propostas por Nevile (2004), Fox (2013) e Prado (2015). As transcrições foram inseridas na ferramenta computacional Wordsmith Tools (Scott, 2016). Essa ferramenta possibilita a realização de um tratamento dos dados linguísticos, que permite, como resultados, a realização de vários tipos de análises, por meio do processamento de um grande volume de informações, eletronicamente, de modo uniforme, preciso, rápido e científico. A seleção desse software foi feita a partir do estudo das obras de Berber Sardinha (2004, 2006), Tagnin e Vale (2008) e Tagnin e Viana (2011). Neste trabalho, daremos ênfase à metodologia empregada nas transcrições e seus desafios, pois a comunicação piloto-controlador tem muitas especificidades de difícil compreensão para quem não está inserido nessa atividade ou mesmo no contexto do momento da fala. Por isso, acreditamos que a pesquisa aqui exposta também apresenta uma contribuição aos estudos sobre transcrição de *corpus* oral, tema ainda carente de mais estudos, principalmente no Brasil. O *corpus* SISCEAB contém um total de aproximadamente 20 horas de gravações efetivamente utilizadas e 210.969 palavras. Foram cobertos diversos temas de rotina, não rotina e emergência, o que confere ao *corpus* representatividade e balanceamento. É possível aplicar filtros para verificar dados relativos somente a um órgão operacional (como APP, por exemplo), ou a uma localidade específica ou região. Também é possível realizar consultas por palavras-chave, temas, colocados, concordâncias. O *corpus* SISCEAB possibilitará a realização de pesquisas tanto sobre a análise e descrição da linguagem piloto-controlador quanto suas aplicações para o ensino-aprendizagem e para a avaliação de proficiência com esse fim específico, notadamente áreas que necessitam de mais estudos, como apontado na última Assembleia da ICAO, realizada em outubro de 2016. Poderá também auxiliar na crítica à escala de proficiência atualmente proposta pela Organização Internacional de Aviação Civil para avaliação desses profissionais, e na proposta de uma escala com base em dados mais científicos. Por fim, espera-se que os benefícios da pesquisa extrapolem o âmbito estritamente linguístico, e possam contribuir para a segurança operacional do tráfego no espaço aéreo brasileiro.

74

COMPILANDO CORPUS PARA APOIO À TRADUÇÃO: TRADUZINDO "*TRADUCCIÓN Y TRADUCTOLOGÍA – INTRODUCCIÓN A LA TRADUCTOLOGÍA*"

Marina Waquil
Universidade Federal do Rio Grande do Sul (UFRGS)
marinawaquil@gmail.com

O presente trabalho tem como base a tradução inédita do espanhol para o português de uma das principais obras da área da tradução: "Traducción y Traductología – In-

troducción a la Traductología", da autora, professora e tradutora espanhola Amparo Hurtado Albir (2001). A obra é a 21ª mais citada nos estudos sobre a tradução, sendo a única publicação editada no século XXI e, além disso, a única em língua espanhola entre as cinquenta mais referenciadas (Franco, 2010). O texto de Hurtado Albir tem caráter fundamental e é reconhecidamente popular nos estudos com foco na tradução, sendo extremamente abrangente ao discutir as diversas definições da tradução, suas características e as principais noções de análise. É, também, democrático, já que dá voz aos mais diferentes teóricos e seus pontos de vista e pressupostos. Assim, a partir da 1) realização da tradução deste texto, estabelecemos como objetivos, também, 2) identificar os problemas terminológicos de tradução decorrentes do processo, apresentando, consequentemente, as soluções propostas para os mesmos, e 3) elaborar um glossário bilíngue espanhol/português, composto por termos coletados da obra de Hurtado Albir e seus respectivos equivalentes. Para o cumprimento destas etapas, compreendemos a necessidade de execução de uma fase fundamental para a pesquisa e para a obtenção dos objetivos estipulados, isto é, a preparação de um *corpus* que servisse de guia para 1) encontrar equivalentes em português para unidades do texto original, 2) resolver os problemas decorrentes da tradução e 3) para validar as unidades terminológicas candidatas a entradas no glossário proposto. Este *corpus*, sua elaboração e posterior utilização, representa, em nosso entendimento, uma etapa fundamental dentro da pesquisa, já que trabalhamos em um contexto de linguagem especializada, a da área da tradução. Desta forma, a tradução realizada não versa sobre um texto comum, mas sobre um texto que tem um caráter especializado (evidenciado segundo os critérios de Cabré, 2001) e, por isso, apresenta características próprias da terminologia da área. Deste modo, buscando mediar com adequação e precisão uma comunicação especializada, acreditamos ser fundamental a busca e confirmação por equivalentes terminológicos e, por isso, procedemos à compilação de um *corpus* de referência da área da tradução no qual pudéssemos realizar esta etapa. Portanto, para sua compilação baseamo-nos na proposta de um dos grandes pesquisadores em Linguística de Corpus no Brasil, Berber Sardinha, que entende que um *corpus* é um conjunto de dados "construído a partir de um desenho explícito, com objetivos específicos" (Berber Sardinha, 2000, p. 335). Para o autor, existem quatro pré-requisitos básicos para a compilação de um *corpus*, os quais buscamos seguir na presente pesquisa. Em primeiro lugar, a autenticidade do conteúdo, isto é, para a formação de um *corpus*, os textos que o compõem devem ser autênticos, tendo sido escritos em linguagem natural e sem que tenham sido criados com o propósito de serem objetos de estudo de pesquisa linguística. Seguindo esse critério, selecionamos artigos de revistas e periódicos brasileiros de destaque no campo da tradução que são autênticos, escritos em linguagem natural e cujo propósito é a difusão do conhecimento especializado (Cadernos de Literatura e Tradução – USP; Cadernos de Tradução – UFSC; Revista Traduzires – UnB; Tradterm – USP; Tradução e comunicação – Anhanguera Educacional; Tradução em Revista- PUC/RIO; Translatio – UFRGS). Ainda em relação à autenticidade está o segundo critério, em referência aos autores dos textos, de modo que, com exceção de alguns casos, como o exemplo dos *corpora* de aprendizes, os textos selecionados para a compilação de um *corpus* devem ter sido escritos por falantes nativos da língua na qual estão escritos. Para o *corpus* deste trabalho selecionamos textos escritos estritamente por pesquisadores falantes nativos de português brasileiro. O terceiro requisito é que a atividade se baseie em uma seleção criteriosa, isto é, que os textos para a

formação de um *corpus* sejam escolhidos de forma que este tenha alguma característica, a partir de regras especificadas pelo(s) pesquisador(es). Seguinte esse critério, selecionamos textos do gênero artigo acadêmico publicados em revistas e periódicos de universidades brasileiras especializados no campo da tradução, já que nosso objetivo era ter um *corpus* de textos compostos pela linguagem que o caracteriza. Finalmente, compõe o quadro de requisitos a representatividade que, no entanto, é também vista como problemática por Berber Sardinha em função da dificuldade em identificar critérios objetivos para defini-la; este pré-requisito deve servir para que o *corpus* seja representativo de uma língua, variedade, campo especializado, etc. No caso deste trabalho, selecionamos um *corpus* de 346 textos, e quase 2,4 milhões de palavras, cuja extensão e fiabilidade deram conta da pesquisa a ser realizada e dos objetivos estipulados para a sua utilização. Com este *corpus* compilado, além de utilizá-lo durante o processo de tradução, buscando e/ou confirmando equivalentes, procedemos à análise terminológica que, segundo Tebé e Cabré (2004, p. 221, tradução nossa[1]), é fundamental já que "assim, o tradutor pode visualizar com algum grau de certeza os tipos de problema que terá que enfrentar e de quais recursos precisará". Dividindo esta etapa em três, primeiro, analisamos o *corpus* de tradução, o texto de "Traducción y Traductología"; logo, o *corpus* de textos compilados sobre tradução em português; e, por fim, comparamos os dados terminológicos obtidos de ambos os *corpora*, com o objetivo de adiantar possíveis questões problemáticas para a tradução. Durante esta análise, alguns dos passos metodológicos aplicados foram a elaboração de *stoplists* em espanhol e em português, com base nas produzidas pelo grupo Termisul, a inserção das *stoplists* e dos *corpora* de textos na ferramenta Wordsmith Tools e a análise dos resultados, observando com maior detalhe as 150 unidades mais frequentes individualmente e comparativamente. A partir desta análise, chegamos a algumas conclusões norteadoras do seguimento do trabalho: por um lado observamos uma clara semelhança na estruturação linguística do conhecimento da área da Tradutologia entre os dois idiomas, com um número extremamente significativo de termos em comum nas duas *wordlists*. A simples análise das palavras mais frequentes de ambas é suficiente para demonstrar esta semelhança, posto que as seis unidades mais frequentes de cada idioma, por exemplo, são praticamente as mesmas: traducción/tradução, texto/texto, traductor/tradutor, lengua/língua, original/original com variação apenas na ordem e no número em texto/textos, no espanhol, e tradutor/tradutores, no português. No entanto, embora este resultado indique uma correspondência significativa entre a linguagem especializada da Tradutologia e a estruturação terminológica da mesma em ambos os idiomas, por outro lado, algumas diferenças significativas também foram constatadas, já que termos chave da área ocupam posições extremamente distantes em ambas as listas. O termo equivalência, por exemplo, representa uma das principais e mais discutidas noções e, embora na *wordlist* em espanhol tenha frequência alta, ocupando a 18ª posição, com um total de 342 aparições, na lista extraída do *corpus* em português, encontramos o termo apenas na 426ª posição, com 280 aparições em um conjunto de textos de aproximadamente 2,4 milhões de palavras. Conclusivamente, a análise terminológica prévia à tradução demonstrou que embora existam semelhanças

[1] No original: Así, el traductor puede visualizar con una cierta seguridad qué tipo de problemas deberá resolver, y qué tipo de recursos necesitará tener a mano, y en suma calcular qué esfuerzo le va a representar el trabajo de traducción y quizá, de forma orientativa, en qué tareas se repartirá este trabajo.

evidentes no interesse temático do conhecimento especializado da Tradutologia em ambas as línguas, os enfoques diferem bastante, o que se evidencia na frequência de uso dos termos. Assim, com base no *corpus* compilado, na análise feita no mesmo e na sua comparação com o texto original, pudemos observar a linguagem da área da tradução, em espanhol e em português, em contextos reais de utilização, entendendo sua estruturação e antevendo questões relevantes para a realização da tradução. Neste trabalho, nosso objetivo é apresentar as etapas de construção do *corpus*, as ferramentas utilizadas e discutir os resultados a que chegamos.

79

READ BD! (THB SIMPLES!)

Heloísa Orsi Koch Delgado, Débora Montenegro Pasin e Carmen Vernetti
Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS),
Instituto de Prevenção e Pesquisa em Álcool e Drogas e outras Dependências (IPPAD)
instituto.ippad@gmail.com, heloisa.delgado@pucrs.br, debora.pasin@acad.pucrs.br, carver@terra.com.br

O estudo proposto trata da simplificação da linguagem psiquiátrica, a do Transtorno do Humor Bipolar (THB), para a comunidade com pouca alfabetização na saúde. A ideia é preencher a lacuna existente entre os profissionais da saúde e seus pacientes através de um guia de bolso intitulado *Read BD!* (em inglês) e THB Fácil! (em português do Brasil). O material explicativo conterá linguagem simples para descrever esta doença de difícil entendimento, resultado da combinação de abordagens quantitativas e qualitativas. Os índices Flesch e o Coh-Metrix medirão o nível de complexidade do *corpus* através das categorias de narratividade, simplicidade sintática, concretude lexical, coesão referencial e coesão lógica. Estratégias de paráfrase e omissão, dentre outras, serão usadas para alcançar índices de simplificação adequados. Questionários serão aplicados e seus resultados comparados a esses índices para assegurar um texto de fácil compreensão.

82

LEITURA DE ARTIGOS CIENTÍFICOS ANGLÓFONOS NA EDUCAÇÃO PROFISSIONAL: *CORPORA*, LÉXICO ESPECIALIZADO E ENSINO

Andrea Jessica Monzón
Instituto Federal do Rio Grande do Sul (IFRS)/Universidade Federal do Rio Grande do Sul (UFRGS)
andrea.monzon@feliz.ifrs.edu.br

Na formação acadêmico-profissional de alunos de cursos técnicos e tecnológicos, existe a demanda pela leitura de artigos científicos em inglês. Os objetivos principais deste estudo foram investigar o léxico especializado anglófono na Computação e observar como os alunos dessa área lidam com as idiossincrasias terminológicas e discursivas na leitura de LE. Para tanto, os dados foram coletados e analisados através da compilação de *corpora*, dois experimentos com aprendizes-alvo e entrevistas com educadores. Os

resultados apontaram para a importância da observação das necessidades linguísticas, pedagógicas e profissionais dos alunos aliada à investigação de gêneros textuais especializados para a construção de estratégias didáticas relevantes.

88

LC E ANÁLISE DO DISCURSO COM ENFOQUE NO ESTILO DO TEXTO LITERÁRIO TRADUZIDO: ITENS LEXICAIS ESTRANGEIROS EM *THINGS FALL APART*

Cristina Lazzerini
Universidade Federal de Minas Gerais (UFMG)
crislazzerini@gmail.com

Este artigo enfatiza a relevância das ferramentas e métodos da Linguística de Corpus para a análise do discurso com enfoque nos traços de estilo de textos literários traduzidos. Baseia-se em um estudo sobre o uso de itens lexicais estrangeiros em um *corpus* paralelo. A LC possibilitou a identificação de padrões de uso dos itens e a integração de abordagens linguísticas e literárias. Os resultados mostram que, embora a estratégia usada pelo autor para a inserção de vocabulário ibo no romance tenha sido seguida pelos tradutores para traduzir os itens, as mudanças de tradução analisadas apontam a construção de significados distintos nas traduções, motivada por uso diferenciado dos itens.

90

"EU QUERO CESÁREA" OU "*JUST CUT IT OUT!*": ANÁLISE CRÍTICA DO DISCURSO DE RELATOS DE PARTO APÓS CESÁREA DE MULHERES BRASILEIRAS E ESTADUNIDENSES À LUZ DA LINGUÍSTICA DE CORPUS

Luciana Carvalho Fonseca
Pontifícia Universidade Católica de São Paulo (FAFICLA/PUCSP)
e Universidade de São Paulo (FFLCH/TRADUSP)
luciana.carvalho@tradjuris.com.br

Este artigo apresenta minha tese de doutoramento em Estudos Linguísticos e Literários em Inglês pela Universidade de São Paulo (Carvalho, 2014 e 2015). O problema social que motivou a pesquisa foi o fato de que, no Brasil, a maioria absoluta das primíparas deseja parto normal logo que engravida, porém, em mais da metade dos casos, os nascimentos são cirúrgicos. O fenômeno da falta de correspondência entre o desejado e o efetivamente alcançado não é exclusividade das mulheres brasileiras, mas ocorre em vários países do ocidente. Por meio da Análise Crítica do Discurso (ACD) de relatos de parto normal após cesárea (relatos de VBAC, do inglês *vaginal birth after c-section*) à luz da Linguística de Corpus (LC), buscamos elucidar o problema social entre a falta de correspondência entre o tipo de experiência desejada e a experiência obtida. O discurso dos relatos de VBAC nos parece ser o discurso ideal para desvelar os elementos dessa falta de correspondência, pois abordam tanto a experiência da cesárea anterior indesejada e, em regra,

mal indicada, como a do parto desejado e alcançado. O recorte teórico-metodológico adotado reúne a ACD (Fairclough, 1989, 1992, 2003; Chouliaraki e Fairclough, 1999); a LC (Stubbs, 1993, 1996; McEnery e Wilson, 1996; McEnery et al., 2003; Tognini-Bonelli, 2001) e a Análise Crítica do Discurso Baseada em Corpus (Baker et al., 2008; Baker, 2013; Baker e McEnery, 2005; Flowerdew, 2014). Na fase da contextualização do problema social, foram analisadas as expressões da mídia sobre o evento 'parto' e 'nascimento'. Verificou-se que a mídia nos apresenta e reapresenta ao longo da nossa vida ideias que desempenham um papel significativo na construção individual e coletiva do que entendemos por nascer e parir. A respectiva análise revelou ideologias e relações de poder que, em seus múltiplos empregos, apresentam o oxímoro – 'parto cesárea' – igualando-o a 'cirurgia', sendo a mulher operada em uma 'sala de parto'. Constatou-se que o 'parto cesárea' é representado como 'opção' da mulher, em uma construção discursiva – e terminológica – que perpetua mitos sobre o 'parto normal' mas evidencia como o impacto da modernidade no discurso publicitário – disfarçado de notícia – empacota experiências de vida de modo a vendê-las como bens de consumo por meio de uma ficção terminológica. Por outro lado, no gênero relatos de parto, serão apresentados os colocados do termo 'parto' – 'parto natural', 'parto humanizado', 'parto domiciliar', etc. – e a vinculação de suas acepções à medicina baseada em evidências e sua contraposição à 'cesárea', tendo em vista que 'parto cesárea' não é um termo empregado no gênero textual dos relatos, apesar de ser um termo que avançou até mesmo sobre os gêneros oficiais do Ministério da Saúde. Após esta imersão nos discursos que circundam os relatos de parto, foi compilado um *corpus* eletrônico em inglês e português. O *corpus* é composto por textos escritos pelas mulheres que passaram pela experiência de VBAC e não inclui textos mediados (entrevistas e relatos escritos por terceiros não foram incluídos). O *Corpus* BRABA (*corpus* eletrônico de relatos de parto de mulheres brasileiras, estadunidenses, britânicas e australianas) se divide, respectivamente, em quatro *subcorpora*: *Corpus* BRA (93 relatos, 250.807 palavras), *Corpus* EUA (101 relatos, 225.736 palavras), *Corpus* UK (97 relatos, 92.197 palavras) e *Corpus* AU (92 relatos, 200.639 palavras). Os primeiros dois *subcorpora Corpus* BRA e *Corpus* EUA foram selecionados para esta pesquisa que pretende investigar como as identidades e a experiência do nascimento são representadas nos relatos de mulheres brasileiras e americanas, e por meio dessa investigação chegar a elementos que elucidem o problema social. O processamento eletrônico valeu-se do programa AntConc 3.4.0w (Anthony, 2012) e das ferramentas da LC (listas de frequência, lista de palavras-chave, linhas de concordância, padrões lexicais, etc.). A análise foi direcionada pelas palavras-chave que correspondem aos sujeitos envolvidos e pelos colocados mais estatisticamente relevantes dessas palavras. No *Corpus* BRA, foram analisadas: eu (colocados: desisto, renasci, mamava); bebê (encaixado, morrer/morresse, sexo, batimentos, alto); marido (companheiro, apoiou, cortou); doula (amada, obstetriz, querida, presença); médico (fofa/fofinha, mudar/mudei, cesarista, ginecologista, humanizada); anestesista; enfermeira (obstétrica/obstetra, cadê, soro, chamar); parteira (liguei/ ligar, doula, casa); obstetriz (doula, toque). No *Corpus* EUA: *I* (*wish*, *protested*, *lamented*); *baby* (*pound*, *girl*, *boy*); *midwife* (*certified*, *asst/assistant*, *student*, *assist*); *doula* (*hired*, *friend*, *called*); *nurse* (*practitioner*, *tells*, *triage*); *doctor* (*office*, *seen*, *comes*); *anesthesiologist*; *husband* (*poor*, *run*, *children*). A análise permitiu que fosse elucidado o problema social em ambas as sociedades e fossem reveladas diferenças discursivas e culturais. A falta de correspondência entre a experiência desejada e a alcançada é representada como tendo sido

causada pela sucessão de eventos distintos. Contudo, em ambos os *corpora*, as experiências são representadas e a autoidentidade e as identidades construídas discursivamente sob a égide dos traços da modernidade, marcadamente, em relação à reflexividade exercida ideologicamente. Porém, a reflexividade é operada não só como um modo de sustentar, mas principalmente como forma de transformar as relações de dominação.

91

A CORPUS LINGUISTICS STUDY ON BRAZILIAN STUDENTS' ACADEMIC VOCABULARY PROFILE VOCABULARY PROFILE

Larissa Goulart da Silva e Ana Paula Seixas Vial
University of Warwick, Universidade Federal do Rio Grande do Sul (UFRGS)
l.goulart-da-silva@warwick.ac.uk, ana.vial@ufrgs.br

The aim of this study is to explore the use of academic vocabulary in a corpus by Brazilian students by answering a) what is the vocabulary profile of assignments written by Brazilian students?; and b) how does it compare to the vocabulary profile of other academic corpora *presented in the literature review? To do so, a corpus of assignments written by Brazilian students was compiled and analysed using Range and* Sketch Engine*. The results show that the corpus by Brazilian students presents the same coverage of academic words as other academic* corpora*.*

94

FERRAMENTA DE AUXÍLIO À PRODUÇÃO DE REDAÇÕES ESCOLARES EM LÍNGUA PORTUGUESA: PASSOS PARA A CLASSIFICAÇÃO DE GRUPOS TEXTUAIS

Aline Evers e Maria José Bocorny Finatto
Universidade Federal do Rio Grande do Sul (UFRGS)
aline.evers@gmail.com, mariafinatto@gmail.com

Este trabalho apresenta os resultados iniciais de uma pesquisa de doutorado em andamento sobre Avaliação Automática de Textos (ATT) em português brasileiro junto ao PPGLetras/UFRGS. São apresentados três estudos-piloto realizados a partir de um *corpus* de redações escolares. Padrões lexicais e coesivos são descritos para dois conjuntos de textos separados de acordo com seus desempenhos em um exame (alto e baixo desempenho). A descrição dos estudos-piloto e dos resultados obtidos servirão para a elaboração de uma ferramenta de auxílio à produção de redações escolares (apoios Capes e CNPq).

99

PROFICIÊNCIA ESCRITA EM LÍNGUA INGLESA PARA A PRODUÇÃO DE *ABSTRACTS* EM CIÊNCIAS DA SAÚDE: UM ESTUDO DE *CORPUS*

Ana Luiza Pires de Freitas e Maria José Bocorny Finatto
Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA),
Universidade Federal do Rio Grande do Sul (UFRGS)
alf@via-rs.net, mariafinatto@gmail.com

Este estudo investiga a proficiência de autores brasileiros na redação de *abstracts* em Ciências da Saúde. A investigação organizou, descreveu e analisou um *corpus* de 180.170 palavras em Medicina, Nutrição e Farmácia. Pacotes lexicais foram eleitos como unidade analítica. A partir das métricas de frequência e variação, os segmentos brasileiro e internacional apontaram distinções nas formas de construir uma narrativa científica. O subcorpus brasileiro apresentou maior repetição de associações de palavras e maior emprego de pacotes lexicais para expressar finalidade e registrar a realização do trabalho acadêmico. Na conclusão, reporta-se a criação de um ambiente virtual de aprendizagem em desenvolvimento a partir da pesquisa.

100

UMA ABORDAGEM PARA IDENTIFICAÇÃO DE PERIÓDICOS ESPECIALIZADOS NO CAMPO DAS HUMANIDADES DIGITAIS

Francisco L. da Cunha e Gyzely S. Lima
Instituto Federal de Educação Ciência e Tecnologia do Triângulo Mineiro – Campus Uberlândia Centro
franciscolopesdacunha@gmail.com, gyzely@iftm.edu.br

Este artigo apresenta os resultados iniciais de um projeto de pesquisa em andamento (PIBIC-CNPq), cujo objetivo é investigar o uso de métodos digitais quantitativos e computacionais que apóiam a prática de pesquisadores no campo das Humanidades Digitais. Através do uso da metodologia de Mapeamento Sistemático, pretendemos fazer um levantamento da literatura sobre pesquisa relacionada às aplicações de métodos digitais na análise de textos literários nas Humanidades Digitais. Neste estágio de pesquisa, apresentamos uma lista de 15 periódicos representativos do campo de Humanidades Digitais que foram identificados pela análise de 26 listas de bibliografias recomendadas publicadas por especialistas.

DICIONÁRIO OLÍMPICO: DESAFIOS E ESTRATÉGIAS PARA O USO DE *CORPORA* NA CONSTRUÇÃO DE UM DICIONÁRIO TEMÁTICO

Rove Chishman e Diego Spader de Souza
Universidade do Vale do Rio dos Sinos (Unisinos)
rove@unisinos.br, dspadersouza@gmail.br

A presente comunicação tem o objetivo de discutir e apresentar os desafios e as consequências do uso de *corpora* textuais na construção do Dicionário Olímpico, desenvolvido e lançado em 2016 pelo grupo de pesquisa Semântica e Tecnologia (SemanTec), da Universidade do Vale do Rio dos Sinos (Unisinos), sob a coordenação da Profa. Dra. Rove Chishman. O Dicionário Olímpico (disponível para consulta em www.dicionarioolimpico.com.br) é um recurso lexicográfico bilíngue (português-inglês) do léxico dos 42 esportes olímpicos. O dicionário está baseado no arcabouço teórico-metodológico da teoria da Semântica de *Frames*, desenvolvida por Charles J. Fillmore (1982; 1985). A Semântica de *Frames*, sendo uma das principais hipóteses da Linguística Cognitiva para a descrição do significado (Salomão, 2013), postula que, ao compreendermos o significado de uma dada unidade ou expressão linguística, estamos, simultaneamente, acessando, em nossa memória de longo prazo, uma estrutura conceptual que armazena informações adquiridas através da experiência; essa estrutura se denomina *frame*. Desse modo, entender o significado de garçom, por exemplo, é o mesmo que entender o que é um *frame* de RESTAURANTE (Croft e Cruse, 2004). No Dicionário Olímpico, isso se traduz em uma forma dupla de consulta à informação lexical: além de uma lista de verbetes organizada alfabeticamente, tem-se, também, uma lista de *frames* (que são chamados de cenários na interface do dicionário). Cada unidade lexical está relacionada a um *frame*, de forma que a descrição do significado está vinculada a uma cena maior, que contextualiza o uso da palavra ou expressão. Todas as unidades lexicais e as informações que possibilitaram a definição dos cenários foram extraídas de *corpora* textuais. O Dicionário Olímpico surgiu como uma extensão de um projeto anterior, que resultou no Field – Dicionário de Expressões de Futebol (Chishman, 2014, disponível em www.dicionariofield.com.br), recurso trilíngue (português-inglês-espanhol) da linguagem do futebol, também baseado na Semântica de *Frames*. O desenvolvimento do Field foi marcado por uma extensa e criteriosa metodologia embasada na Linguística de Corpus, tendo sido utilizado o software *Sketch Engine* (Kilgarriff et al., 2014, disponível em www.sketchengine.co.uk) em praticamente todas as fases de extração e análise da informação linguística. De início, pensou-se em replicar a mesma metodologia na construção do Dicionário Olímpico, o que logo mostrou-se desafiador e problemático. O primeiro problema encontrado foi a escassez de material para compor os *corpora*, tanto em português quanto em inglês. No caso do Field, foram construídos três *corpora* de cerca de um milhão de palavras cada a partir de textos do gênero *match report* (cf. Schmidt, 2008), que narra os acontecimentos de uma dada partida esportiva. No contexto do Dicionário Olímpico, percebeu-se que a pouca popularidade da grande maioria dos esportes resultava em uma baixa disponibilidade de material. Nesse sentido, recorreu-se a manuais de regras oficiais e textos didáticos disponibilizados online por confederações, treinadores, esportistas e até

mesmo professores de Educação Física. Esses *corpora*, montados a partir desse tipo de material, possibilitaram, antes de mais nada, a familiarização com os esportes. A partir da leitura manual desses textos, portanto, foi possível fazer não apenas a coleta das unidades lexicais de cada modalidade, mas também compreender a dinâmica dos esportes, o que auxiliou no processo de definição dos *frames* – a leitura dos manuais ainda permitiu a construção de mapas conceituais, que organizaram informações enciclopédicas de organização dos esportes. O *corpus* em inglês foi utilizado para a extração de equivalentes de tradução e de exemplos em língua inglesa. O problema da escassez de material no gênero *match report* também se apresentou nesse caso, e o material em língua estrangeira também foi coletado a partir de confederações e profissionais da área esportiva. A validação dos equivalentes de tradução e a busca por mais exemplos, que se fez necessária no caso de esportes de baixíssima popularidade, como o tiro com arco, levou à utilização da *web* como *corpus*, de forma que foram feitas pesquisas em *sites* de busca, como o Google, no intuito de confirmar algumas informações. Nesse sentido, a metodologia da *web* como *corpus* (cf. Gatto, 2012) também foi um meio utilizado para suprir a falta de *corpora* mais robustos, que pudessem ser processados em softwares como o *Sketch Engine*. Vale salientar, também, que o projeto contou com atletas profissionais, professores e técnicos que atuaram como consultores, avaliando e corroborando as informações coletadas. Ainda que não tenha sido possível fazer uso da mesma metodologia do Field – Dicionário de Expressões do Futebol, as estratégias encontradas pelo grupo de pesquisa SemanTec possibilitaram que, mesmo assim, todas as informações presentes no Dicionário Olímpico tenham sido extraídas de textos naturalísticos. Além disso, os gêneros utilizados – regras oficiais e manuais didáticos – permitiram um estudo profundo das modalidades esportivas, dos conceitos que os compõem e da dinâmica que existe entre todas as partes de cada esporte. Tendo isso em vista, entendemos que, apesar de a metodologia utilizada não apresentar, de forma alguma, a sofisticação das tendências mais atuais da Linguística de Corpus, permitiu que todo o estudo linguístico e desenvolvimento do dicionário fosse feito a partir de dados naturalísticos, extraídos do mundo real, o que se relaciona diretamente com o posicionamento filosófico-epistemológico da Semântica de *Frames* (e, de forma geral, da Linguística Cognitiva): a linguagem está, ao mesmo tempo, no mundo e na mente dos falantes.

105

ESTUDO DE CASO BASEADO EM *CORPUS* SOBRE O PERFIL ESTILÍSTICO DE UM TRADUTOR LITERÁRIO

Carolina Barcellos
Departamento de Línguas Estrangeiras e Tradução/ Instituto de Letras, Universidade de Brasília (UnB) Brasília, DF – Brasil
carolinabarcellos@unb.br

Considerando os estudos de estilo da tradução, no contexto dos Estudos da Tradução Baseados em Corpus (Baker, 2000; Saldanha, 2011; Walder, 2013), este trabalho apresenta um estudo dos traços estilísticos de um tradutor literário sob a perspectiva da convencionalidade e da análise de mudanças na tradução (*shifts*). O objetivo foi verificar a

existência de um conjunto de padrões de escolhas do tradutor a respeito da convencionalidade (Baker, 2007) em língua portuguesa brasileira que pudesse ser encontrado tanto nos seus textos traduzidos quanto nos seus textos não traduzidos, e as consequências dessas escolhas para a construção de significado nos textos traduzidos. Foram investigadas também as mudanças na tradução identificadas no contexto de ocorrência dos elementos de convencionalidade analisados a fim de se obter mais informações sobre as preferências linguísticas do tradutor (Pekkanen, 2010; Blauth 2015). Saldanha (2011) parte do trabalho pioneiro de Baker (2000) sobre estilo da tradução em *corpus* comparável, propondo uma abordagem combinada para investigação do estilo do tradutor e estilo do texto traduzido em *corpus* paralelo, a qual permitiria diferenciar marcas do estilo do tradutor das interferências do estilo dos autores dos textos-fonte. Baker (2007) inaugura o estudo da convencionalidade associado ao estilo do tradutor, valendo-se de um *corpus* comparável e aponta a necessidade de aprofundar sua investigação em um *corpus* paralelo. Munday (2008), que investiga o estilo do tradutor de uma perspectiva distinta dos Estudos da Tradução baseados em Corpus, encontra indícios de padronização da voz dos autores quando traduzidos por um mesmo tradutor e associa o uso da convencionalidade à criatividade. Walder (2013) parte dos resultados desses trabalhos e investiga as escolhas de um tradutor através do cotejamento entre um *corpus* de textos traduzidos e de textos não traduzidos desse tradutor/autor. No contexto do GRANT, Magalhães e Barcellos (2015) e Magalhães e Blauth (2015) enfocaram o estilo do tradutor em *corpus* paralelo sem tratar especificamente da convencionalidade em textos traduzidos ou de sua relação com a criatividade do tradutor. A presente pesquisa se propõe a preencher as lacunas encontradas nos trabalhos citados, investigando a convencionalidade relacionada explicitamente à criatividade, da perspectiva do estilo do tradutor e do texto traduzido em um *corpus* de estudo que engloba textos traduzidos e não traduzidos de um mesmo tradutor/autor. Para esta pesquisa, foram compilados três *corpora*: 1) um *corpus* de textos traduzidos para a língua portuguesa brasileira por Paulo Henriques Britto, 2) um *corpus* de textos não traduzidos escritos em língua portuguesa brasileira por Paulo Henriques Britto, e 3) um *corpus* paralelo de contos escritos em inglês americano, dos autores Philip Roth, John Updike, e Jhumpa Lahiri, e suas traduções para o português brasileiro, de Paulo Henriques Britto. Foram utilizados ainda dois *corpora* de consulta (COMPARA e ESTRA) para obter frequências de referência, em língua portuguesa brasileira, quanto ao emprego dos elementos de convencionalidade investigados nos textos de Britto. Na primeira fase desta pesquisa, foram analisados os dados estatísticos do *corpus* de estudo obtidos com o auxílio do programa WordSmith Tools 6.0 e foi feita a contabilização de elementos relacionados à convencionalidade (Britto, 2012) em língua portuguesa em vários níveis (morfema, palavra, grupo e oração). Na segunda fase, foram investigadas as mudanças na tradução (Catford, [1965] 1978) no cotexto dos elementos de convencionalidade destacados na primeira fase. A metodologia de investigação adotada seguiu os preceitos postulados pela Linguística de Corpus e compreendeu compilação, preparação, alinhamento e etiquetagem dos textos para posterior análise com o auxílio do programa WordSmith Tools 6.0. Os resultados alcançados apontaram que o estudo da convencionalidade está relacionado ao uso criativo da linguagem pelo tradutor e, ainda, que sua investigação pode fornecer subsídios tanto para a elaboração do perfil estilístico de tradutores individuais quanto para a identificação do estilo do texto traduzido. As características singulares do *corpus* de estudo compilado para a presente pesqui-

sa permitiram separar a voz do tradutor da voz dos autores dos textos-fonte e possibilitaram também identificar preferências linguísticas desse tradutor presentes tanto nos seus textos traduzidos quanto nos seus textos não traduzidos. Os resultados indicaram que Britto fez um conjunto de escolhas distinto para os textos traduzidos de cada autor do *corpus*, sendo influenciado, embora não de maneira determinante, pelo estilo dos autores dos textos-fonte. Em geral, as escolhas linguísticas de Britto em relação ao emprego de expressões convencionais aumentaram o grau de coloquialidade dos textos traduzidos quando comparados aos seus respectivos textos-fonte. O padrão de escolhas identificado nos seus textos não traduzidos dialoga com o padrão de escolhas identificado nos seus textos traduzidos, em particular com o conjunto de textos traduzidos a partir da obra de Philip Roth. A mudança na tradução mais frequente foi o acréscimo (subcategoria da amplificação). Essas instâncias de acréscimo não estavam diretamente relacionadas à explicitação, mas sim à inserção de expressões convencionais típicas da língua portuguesa, sobretudo quando não havia motivação clara para isso nos textos-fonte. Britto fez uso ainda do recurso de sanitização, apagando algumas referências culturais presentes nos textos-fonte. Entretanto, o número de ocorrências de uso criativo da linguagem realizadas pelo tradutor superou o uso de sanitização, corroborando os dados de Munday (2008) e refutando, até certo ponto, o que foi postulado por Baker (1999, 2000). O presente trabalho oferece, dessa forma, informações sobre o papel das preferências linguísticas do tradutor nas escolhas que ele faz para a realização de significado nos textos traduzidos e, portanto, oferece informações também sobre o mundo que o tradutor escolheu recriar para o leitor da língua-alvo, discutindo as fronteiras entre reconstrução de significados, naturalidade, fluência e precisão no texto traduzido.

Resumos de trabalhos em pôster

6

REARRANGEMENT AND CREATION OF NEW CORPORA FOR UPDATE AND COMPRESSIVE SUMMARIZATION TASKS FOR PORTUGUESE LANGUAGE

Fernando Antônio Asevedo Nóbrega e Thiago Alexandre Salgueiro Pardo
fasevedo@icmc.usp.br, taspardo@icmc.usp.br

We introduce in this paper two corpora *for new summarization tasks for Portuguese language. We report (i) the rearrangement of an existent multi-document summarization corpus in order to be used for update summarization and (ii) the creation of a corpus for training compressive summarization methods.*

7

PARA A SEGMENTAÇÃO AUTOMÁTICA DE FRONTEIRA NA FALA ESPONTÂNEA A PARTIR DE PARÂMETROS PROSÓDICOS

Bárbara Falcão, Plínio A. Barbosa e Tommaso Raso
Departamento de Linguística, Faculdade de Letras – Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil; Departamento de Linguística, Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, UNICAMP, Campinas, Brasil
barbaraheloha@gmail.com, pabarbosa.unicampbr@gmail.com

O trabalho utiliza dados de *corpora* de fala espontânea para realização de um *script* do Praat para detecção automática ou semiautomática de fronteiras prosódicas. O *script* contribuirá ao melhor entendimento da segmentação da fala, porque utiliza como critério de referência para a detecção automática dessas fronteiras os parâmetros acústicos associados à percepção humana. O *script* também auxiliará na compilação de *corpora* de fala espontânea, porque poderá tornar o processo de segmentação da fala mais rápido, poupando-se simultaneamente tempo e esforços humanos, o que pode ser visto como uma contribuição para a linguística de *corpus* em geral.

12

THE USE OF CORPUS LINGUISTICS IN TEACHING ENGLISH FOR ACADEMIC PURPOSES

Fabiana Klein
Instituto de Letras – Universidade Federal do Rio Grande do Sul (UFRGS)
fabimklein@gmail.com

This paper aims to analyze the effectiveness of different corpus-based activities related to the process of writing abstracts for academic papers by basic level English students in a Brazilian public university. The study was conducted during the year of 2016 with two basic level EAP groups from the project called Idiomas sem Fronteiras (Languages without Borders). All the

courses from the project are offered for free to the academic community and aim at helping students of English as a Foreign Language (EFL) in their internationalization process, in order to spread the Brazilian research throughout the academic world as a whole. This was done, firstly, by describing the development of different activities using corpora *(compiled with academic papers from the students' areas of study in the online widely available text analysis tool called Voyant (https://voyant-tools.org/) and the performance of two different groups of students who would taking the same course: "English for Academic Purposes: Abstract and Article Review". The first group of students who had their writing productions analyzed did not use corpus-based materials while writing and rewriting their abstracts. They constituted the control group. Their data was compared with the performance of a group students who have used lexical bundles and* corpora *consultation while writing and rewriting their abstracts, during the same amount of time and same structure of classes. All the students who wrote abstracts in this process were master's students at UFRGS (Federal University of Rio Grande do Sul) and were placed as A2 level students in the Common European Framework of Reference for Languages. Although taking in consideration that the results of the study may be influenced by the small number of participants and the cognitive and personal differences between students, the results show that using* corpora *can increase the quality of their academic writing without necessarily repressing their creativity. In addition, it is stated that using these resources in the classroom requires the teacher to adapt activities in order to stimulate the students' autonomy (when searching for information on the academic writing from their fields) and creativity (when writing their own papers). It is important to consider that the traditional methods of learning and teaching languages usually follow the model of studying grammar rules and seeing examples only after that. As corpus-based materials and activities demand the student to deal first with the current examples of how to convey some type of information, the teacher must be willing to propose discussions and activities that aim to sharpen the group's abilities when working with written comprehension and production. This can be done even with non-academic texts, for a start, but it is important to engage the group in every class attempts on improving writing creatively. Coupled with these considerations, the study allowed the teacher to notice that the students who used* corpora *consultation had fewer difficulties in using the correct prepositions and collocations in sentences, since using* corpora *allowed them to analyze the main phrases and combination of words through their texts. This study may demand, in the future, more specific research on how to create and adapt new corpus-based materials and methods and how to teach students to compile and use* corpora *independently.*

14

OS VERBOS DO DISCURSO ACADÊMICO ESCRITO EM UM *CORPUS* DE APRENDIZES DE INGLÊS

Annallena de Souza Guedes
Universidade Federal de Minas Gerais (UFMG)/IFBA
souzaanna2014@gmail.com

Estudos com *corpora* de aprendizes (Hunston, 2002) têm contribuído para a pesquisa na área de ensino e aprendizagem de língua adicional, principalmente, pela utili-

zação de abordagens de análise linguística, tal como a Análise Contrastiva de Interlíngua (ACI) (Granger, 1996) que, dentre outros objetivos, busca comparar produções escritas de aprendizes a *corpora* de falantes nativos. Considerando os princípios da ACI, essa pesquisa de doutoramento tem como propósito analisar a frequência e os padrões de uso do que chamamos de 'verbos do discurso acadêmico escrito' (VDAEs) em inglês, ou seja, verbos mais típicos no contexto das produções escritas da academia, no gênero redação argumentativa. Nesse sentido, analisamos um *corpus* de aprendizes de inglês, o CorIFA (*Corpus* do Inglês para Fins Acadêmicos), compilado na Universidade Federal de Minas Gerais desde 2013 e, utilizamos como *corpus* de referência, o MICUSP (*Michigan Corpus of Upper-Level Student Papers*), compilado na Universidade de Michigan, nos Estados Unidos. Decorrente desse objetivo central, este estudo se propõe, ainda, a analisar as colocações mais frequentes que coocorrem com tais verbos. Como nos lembra Sinclair (1991), o significado de uma palavra, bem como seus padrões léxico-gramaticais, são interrelacionados, o que nos leva a perceber que o contexto de uso no qual a palavra se insere nos ajuda a compreender as características distintivas entre as escolhas linguísticas dos aprendizes e de usuários mais proficientes da língua. Para a obtenção dos dados, tomamos como base os seguintes procedimentos metodológicos: primeiro, geramos a lista de palavras do CorIFA, através do programa *AntConc* 3.2.4 e usamos o anotador *Claws* 7, a fim de identificarmos os verbos mais frequentes do *corpus*, considerando o ponto de corte de frequência de 20 por 100 mil palavras (Cortes, 2008). Em seguida, buscamos quais verbos, tidos como mais frequentes no CorIFA, compõem a *New Academic Vocabulary List* (AVL) (Gardner e Davies, 2013). Feito isso, comparamos a frequência e o uso desses verbos, bem como suas colocações com os dados encontrados nas redações argumentativas do MICUSP. Os resultados parciais nos mostram que há pouca variação lexical e gramatical nos textos do CorIFA, principalmente no que diz respeito ao uso dos verbos. Ainda, tendo em vista as análises parciais de uso dos VDAEs e suas colocações pelos aprendizes, verificamos que esta pesquisa confirma estudos anteriores (Granger, 1998; Nesselhauf, 2004), nos quais aprendizes de inglês como língua adicional tendem a usar com pouca frequência os verbos mais típicos do discurso acadêmico de falantes nativos. Podemos citar, como exemplo, '*develop*' (143,13) e '*improve*'(135,56), como os verbos de maior frequência normalizada no CorIFA. Além disso, observamos que os aprendizes têm dificuldades ao fazer as combinações fraseológicas com esses verbos, muitas vezes, dando preferência àqueles de alta frequência. Observando o uso do verbo *depend*, encontramos muitas ocorrências da colocação '*depend of*'(em detrimento de '*depend on*'), no CorIFA, ao passo que não encontramos nenhuma ocorrência desta colocação no MICUSP. Esse fenômeno pode se justificar pela influência da L1 do aprendiz na sua escrita e, mais especificamente, nos padrões colocacionais do inglês. Desse modo, podemos dizer que há entre os aprendizes pouca familiaridade com as fraseologias próprias do discurso acadêmico escrito, necessárias à produção de textos em contextos universitários. Pensando nos resultados que obtivemos até então, acreditamos que nossa pesquisa possa oferecer importantes contribuições para o ensino, mais particularmente de inglês como língua adicional, na medida em que evidenciam as dificuldades encontradas pelos aprendizes e que são, muitas vezes, desconhecidas pelos seus professores. Assim, propomos a criação de uma lista ou material de consulta sobre os VDAEs com suas colocações mais frequentes, no intuito de auxiliar os estudantes universitários, aprendizes de inglês, nas escolhas lexicais e gramaticais, que sejam mais apropriadas e condizentes com o uso da língua circulante no

discurso acadêmico escrito. Portanto, o presente estudo constitui um passo importante justamente porque aponta vários aspectos que distinguem o *corpus* de aprendizes do *corpus* de usuários proficientes quanto ao uso dos verbos típicos do discurso acadêmico. Ao mostrarmos como esses usos se efetivam na escrita dos aprendizes, estamos em vias de melhor compreender os textos que produzem e, dessa forma, pensar em como podemos auxiliá-los na tarefa de chamar a sua atenção para conhecer e entender como operam os usos lexicais e gramaticais do inglês. Nessa perspectiva, defendemos ser importante investigar os verbos do discurso acadêmico escrito, no sentido de podermos ter subsídios que melhor caracterizem seu uso e, assim, contribuirmos para preencher lacunas quanto aos estudos sobre escrita acadêmica, principalmente no que diz respeito aos textos produzidos em inglês, direcionados a um público específico e motivados por propósitos diversos, no intuito de atender a diferentes demandas acadêmicas.

18

A NOTÍCIA POLICIAL NA MIRA DA LINGUÍSTICA DE CORPUS

Joacyr Oliveira
Universidade de São Paulo (USP)/Universidade Estadual de Campinas (UNICAMP)
joacyr@outlook.com

A pesquisa aqui apresentada é um ensaio para demonstrar a aplicação da Linguística de Corpus como uma ferramenta de facilitação e de estatística em análises propostas pela Análise do Discurso Crítica (ADC). Para tanto, ocuparemo-nos, dentro do gênero jornalístico informativo, de notícias policiais. É notório que a imprensa, escrita ou falada, diligencia ser isenta/imparcial em relação aos fatos que noticia. Até que ponto e grau essa imparcialidade é, de fato, atingida? Como objeto de estudo desta investigação, escolhemos a divulgação do assassinato de Mércia Nakashima, ocorrido em 23 de maio de 2010. O assassino, julgado e condenado em agosto de 2013, foi seu ex-namorado, Mizael Bispo de Souza. A escolha desse crime se deve ao fato de tanto a vítima quanto o homicida terem a mesma profissão: advogado. Há um terceiro personagem: o cúmplice Evandro Bezerra da Silva, ora referido como vigia, ora como vigilante. O *corpus* é composto por 94 textos, sendo 47 artigos do jornal O Estado de São Paulo e 47 artigos do jornal Folha de São Paulo, publicados do dia 23 de maio de 2010 a 27 de agosto de 2010, um mês após a justiça ter pedido a prisão preventiva de Mizael Bispo. O *corpus* totaliza 3.551 *types* e 40.222 *tokens*. A pergunta que norteia a pesquisa é se os textos das matérias publicadas associam os nomes da vítima e do assassino à profissão de advogado e do cúmplice à profissão de vigia da mesma forma. Dados fundamentados na análise do *corpus* demonstram que, na verdade, cada um dos envolvidos é tratado de forma distinta. Iniciamos nossas análises a partir da lista de palavras-chave utilizando a abordagem *corpus-driven*. Os cinco termos com maior índice de chavicidade e frequência são Mizael (613 ocorrências), Mércia (447), advogado (239), Nakashima (185) e advogada 190. A princípio, parece que as profissões são associadas aos envolvidos de forma igual. Entretanto, a análise dos colocados e das linhas de concordância com os termos “advogada” e “advogado” revelaram um dado inesperado. Enquanto “advogada” se coloca com “Mércia” 59 vezes, em todas as outras ocorrências o termo é usado de forma pronominal,

ou seja, substituindo o nome da vítima. Já o termo "advogado" se coloca 37 vezes com "Mizael" e apenas uma vez é usado de forma pronominal. Em todas as outras ocorrências, o termo é usado para se referir aos advogados envolvidos profissionalmente no caso. Nesta conjuntura da pesquisa, podemos concluir que, enquanto a vítima é constantemente associada à sua profissão, o assassino é referido por seu nome ou sobrenome ou como "ex-namorado" (63 ocorrências). Tais aspectos das notícias podem levar o leitor a "desassociar" o assassino da profissão de advogado. Durante nossa exposição, apresentaremos a lista de palavras, a lista de palavras-chave, e as linhas de concordância para "Mizael", "Mércia", "advogada", "advogado", e incluiremos os resultados das pesquisas com o nome e profissão do cúmplice do crime: "Evandro" e "vigia". "Evandro" aparece em 11º lugar com 119 ocorrências e "vigia" + "vigilante" em 23º lugar com 84 ocorrências.

20

GRAMÁTICA LIVRE DE CONTEXTO LEXICALIZADA E A LINGUAGEM RADIOLÓGICA

Cinthyan Renata Sachs C. de Barbosa, Davidson Cury e Edio Roberto Manfio
Departamento de Computação – Universidade Estadual de Londrina (UEL);
Instituto de Informática – Universidade Federal do Espírito Santo (UFES);
Faculdade de Tecnologia Deputado Júlio Julinho Marcondes de Moura (FATEC)
cinthyan@uel.br; dede@inf.ufes.br; prof.ediorobertomanfio@gmail.com.br

Este trabalho apresenta uma nova abordagem à Gramática da Língua Natural para descrição do português através de um sistema de geração de árvores, chamado Gramática Livre de Contexto Lexicalizada (GLCL), a qual foi testada utilizando um subconjunto da linguagem radiológica, onde foi possível validar um diagnóstico médico.

25

O VOCABULÁRIO DE *PRIDE AND PREJUDICE* E SUA TRADUÇÃO PARA O PORTUGUÊS: UMA ANÁLISE LEXICOESTATÍSTICA

Laís Callegaro Fritzen e Luísa Branchi Araújo
Universidade Federal do Rio Grande do Sul (UFRGS)
lais.fritzen@gmail.com; branchi.luisa@gmail.com

Este artigo apresenta a análise lexical comparativa de *Pride and Prejudice*, de Jane Austen, e uma de suas traduções para o português brasileiro. O objetivo foi estudar a complexidade lexical por meio da lexicoestatística, à luz dos estudos da Linguística de Corpus. Fez-se o levantamento do léxico com o AntConc 3.4, em cotejo com os *corpora* de referência – Corpus of Contemporary American English e Projeto PorPopular – e comparou-se os níveis de complexidade lexical com o AntWordprofiler. Constatou-se que a lexicoestatística serve aos estudos de complexidade lexical e textual ao auxiliar o pesquisador a explorar o léxico em seu uso real. Ao propiciar a identificação do público leitor, contribui ainda aos estudos de gênero textual e de tradução.

26

TICAL: UM *CHATBOT* QUE VERSA SOBRE ASSUNTOS LINGUÍSTICOS

Fábio Carlos Moreno, Edio Roberto Manfio e Cinthyan R. Sachs C. de Barbosa
Departamento de Computação – Universidade Estadual de Londrina (UEL);
Faculdade de Tecnologia Deputado Júlio Julinho Marcondes de Moura (FATEC)
fbio_moreno@yahoo.com.br; prof.ediorobertomanfio@gmail.com.br; cinthyan@uel.br

Este artigo aborda a construção de léxicos digitais no âmbito do Processamento de Língua Natural (PLN). Será descrito a seguir um robô de conversação chamado TICAL que versa sobre o ALiB (Atlas Linguístico do Paraná) e tem aumentado seu domínio para o Léxico Histórico do Paraná (LHIsPAR), que também constitui-se como história do Português Brasileiro.

29

VOCABULÁRIOS MONOLÍNGUES DE TERMOS DA TEOLOGIA: ESTUDO CONDUZIDO POR *CORPUS*

Solange Cardoso
Universidade Federal de Uberlândia (UFU)
solangejac@yahoo.com.br

Este trabalho objetiva a elaboração de três propostas de vocabulários terminológicos da área de Teologia que usará a ferramenta do Vocabulário Técnico Online (VoTec) (Fromm, 2007), que pode ser visualizada, gratuitamente, em página da internet (www.pos.voteconline.com.br). As bases teóricas desta pesquisa estão alicerçadas nos conceitos teóricos e metodológicos da Teoria Comunicativa da Terminologia (TCT; Cabré, 1993; 1999); nas concepções de dicionário, vocabulário e glossário de Barbosa (2001); nos estudos de Barros (2004), Krieger e Finatto (2004) e Almeida, Pino e Souza (2007), em relação à elaboração das definições dos verbetes; nos pressupostos teóricos para o trabalho com a metodologia da Linguística de Corpus (Berber Sardinha, 2004; Teixeira, 2008; Scott, 2012) e na operacionalização da plataforma de gestão terminológica VoTec (Fromm, 2007). Os *corpora* textuais, que serviram à extração das Unidades Terminológicas e à elaboração da proposta de vocabulários, foram constituídos de textos escritos e digitalizados utilizados pelos alunos, professores e formadores nas aulas e nos encontros de evangelização da Faculdade Shalom de Ensino Superior (Fases) e da Faculdade Católica de Uberlândia (FCU). Consideramos que este estudo se justifica pelo fato de que os dicionários representam hoje uma ferramenta de grande importância no processo de ensino/aprendizagem das diferentes áreas do conhecimento humano e, nesse sentido, a reflexão sobre a Terminografia e a preocupação com a elaboração de dicionários para aprendizes têm crescido ultimamente. Além disso, este trabalho contribuirá para a criação de vocabulários com verbetes cuja construção das microestruturas será baseada em dados coletados por meio de *corpora* específicos. Assim, as definições (microestruturas) serão construídas unicamente por meio de exemplos tirados dos *corpora* por nós compilados. Procederemos assim porque, raramente, produtos terminográficos

são criados a partir do perfil e das necessidades dos aprendizes de uma área de especialidade em determinada situação comunicativa. Ou seja, raras são as obras elaboradas para um público constituído por aqueles que ainda se encontram em um período de formação, seja ela técnica ou acadêmica. Como ministramos aulas de língua portuguesa em curso de Bacharel em Teologia, temos presenciado a dificuldade de estudantes universitários para a compreensão de textos, cuja área especializada é a Teologia. A dificuldade não é superada ainda que esses estudantes sejam levados a consultar obras lexicográficas do tipo: dicionário secular, dicionário bíblico, chave (ou referência) bíblica, livro de geografia bíblica e ou enciclopédia bíblica. Isso nos remete principalmente à questão da eficiência dessas obras no que se refere ao auxílio das aprendizagens no campo das ciências teológicas (principalmente, de alunos nos anos iniciais dessas aprendizagens). Esta proposta de trabalho pretende, então, ser uma contribuição para a elaboração de dicionário monolíngue de termos da Teologia, tendo como ponto de partida a compilação de *corpora* a partir das bibliografias indicadas nas ementas e outras sugeridas pelos professores de cursos de Teologia. A elaboração deste trabalho se justifica também pelo fato de haver diferentes denominações cristãs, cuja distinção, certamente, fixa-se por meio de conceitos que estabelecem o cerne de formação teológica desenvolvida, particularmente, em cada uma dessas denominações. As etapas metodológicas que seguimos estão assim organizadas: i) elaborar Árvores de Domínio da área de Teologia das instituições Comunidade Shalom e Igreja Católica Apostólica Romana, conforme a orientação teológica das faculdades pesquisadas; ii) selecionar, coletar e organizar os *corpora* de estudo; iii) elencar candidatos a termos que possam ser incorporados aos vocabulários; iv) analisar, contrastivamente, os vocabulários das duas faculdades para a geração do terceiro vocabulário; e v) inserir os dados na plataforma de gestão terminológica do VoTec. Esclarecemos que este trabalho ainda não foi finalizado, apresentamos assim, a organização e análise dos *corpora* de pesquisa: etapas i) e ii). Elaboramos a Árvore de Domínio da grande área da Teologia pesquisando o sítio virtual do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e o sítio virtual da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes). Por meio da consulta a documentos e especialistas das duas faculdades, elaboramos as árvores de domínio dos cursos de Teologia. Iniciamos a organização dos *corpora* compilando livros indicados pelos alunos de uma das faculdades. Esclarecemos que essa indicação teve a devida anuência dos professores. A seguir, fizemos inúmeras incursões a diferentes sítios virtuais. Para essas incursões, usamos as áreas e subáreas (Árvore de domínio) como palavras-chave de busca. A composição do *corpus* reuniu: apostilas elaboradas por professores, livros-texto, manuais de Teologia, entrevistas com estudiosos da área, sermões, encíclicas, etc. Além dos livros de indicação dos alunos, todo o material foi reunido por meio de consultas a sítios devocionais católicos e evangélicos. Consultamos também e coletamos material em diferentes sítios de cursos superior de Teologia da modalidade a distância (EaD). Localizados os textos, esses eram selecionados e inseridos em arquivos do Word para a realização de revisão visto que, no processo de seleção, cópia e colagem dos textos, caso houvesse erros, esses seriam visualizados. Após essa etapa, foram inseridos cabeçalhos contendo as informações mais gerais pertinentes a cada texto: título e autoria (quando localizada) e os arquivos foram salvos em txt. No final do texto, inserimos local (endereço virtual) e data de coleta. Criamos códigos para a nomeação dos arquivos usando a letra inicial de cada palavra do título do texto, seguida das letras iniciais das subáreas e áreas

da Teologia. O título, a subárea e a área foram separados por hífen. Reunimos todos os textos em dois grandes arquivos: no primeiro, os textos da FCU e no segundo, os textos da Fases. Como ferramenta de análise lexical, utilizamos o WordSmith Tool (WST) que permite, dentre outras funções, gerar listas de palavras, palavras-chave e concordâncias. Para gerar as listas de palavras-chave, comparamos a frequência das palavras dos *corpora* de estudo com o *corpus* de referência do Banco de Português (BP). Para a inserção na plataforma do Votec, estabelecemos os critérios de escolha dos candidatos a termos. Assim, i) selecionamos 45 candidatos a termos entre os 100 primeiros itens mais chaves; ii) selecionamos, por relação lógica, o termo contido na definição de outro cuja compreensão é fundamental para o entendimento do termo definido. Assim, tal termo também será definido e será remissivo. Por exemplo, "Deus é a Trindade Santa"; o termo Trindade deverá ser definido e será remissivo ao termo "Deus"; iii) preferencialmente, serão definidos os termos correspondentes entre os dois *corpora* de estudo, pois o objetivo é realizar uma análise contrastiva desses termos; iv) após a aplicação do terceiro critério, não satisfeita a prerrogativa de se encontrar um candidato a termo correspondente, no *corpus* em contraste, foram considerados os termos com alta chavicidade e relevância lógica dentro do campo do conhecimento teológico católico e do conhecimento teológico evangélico. Por exemplo, o termo Eucaristia está presente no CC e não aparece no CE. A escolha desses termos, nesse quarto nível, possibilitará a comparação em segunda instância, ou seja, poderemos detectar/analisar não o contraste do conteúdo da definição dos termos, mas também, a diferença de termos mais ou menos relevantes entre as duas linhas teológicas; v) quando se tratava de um candidato a termo que fosse composto, embora aparecesse isoladamente na lista de palavras-chave e fazia sentido separado, como por exemplo, Jesus (em Jesus Cristo) e Espírito (em Espírito Santo), optamos por definir o termo composto. Para isso, alguns procedimentos foram necessários, a fim de gerar as linhas de concordâncias do termo composto. Passamos então a descrever os procedimentos a serem executados. O primeiro passo foi localizar a palavra principal ou primeira à esquerda do termo composto (Jesus, por exemplo). Em seguida, gerar a linha de concordância daquele termo e salvá-la. O terceiro passo foi abrir a função "*collocates*" e encontrar o "colocado" com o termo composto (Cristo). O WST disponibiliza a quantidade de colocados com o item desejado, clicamos sobre ele para abrir as linhas de concordâncias com o colocado desejado (Jesus Cristo). Assim, pudemos localizar as linhas que tivessem possíveis contextos definitórios e/ou explicativos, que permitissem a elaboração das definições; vi) outro procedimento a ser adotado para a localização de contextos eficientes para a criação da definição foi a utilização da função *clusters* já inserido na tela de concordância do colocado, "em processo de escolha para ser um termo definido". Pretendemos que o produto final seja um instrumento que facilite a compreensão de conteúdos/textos objetos de estudo das aulas ministradas por professores dos anos iniciais de cursos de Teologia. Nosso desafio é, por meio dos textos que constituirão os *corpora* desta pesquisa, elencar os termos e as definições que, provavelmente, constituem a visão do conhecimento acadêmico teológico difundido pela Faculdade Católica de Uberlândia (Vocabulário dos termos da Teologia – Faculdade Católica de Uberlândia) e pela Fases (Vocabulário dos termos da Teologia – Fases). O terceiro dicionário (Vocabulário os termos em contraste) será o resultado do confronto entre termos idênticos utilizados nos dois outros vocabulários.

TERMINOGRAFIA DIDÁTICO-PEDAGÓGICA: METODOLOGIA PARA APOIO À LEITURA EM INGLÊS DE TEXTOS DE ELETROTÉCNICA/ENGENHARIA ELÉTRICA

Sabrina Bonqueves Fadanelli
Universidade Federal do Rio Grande do Sul (UFRGS)/Universidade de Caxias do Sul (UCS)
sbfadane@ucs.br

A motivação para esta pesquisa nasceu no período em que iniciei meu trabalho como docente na disciplina de Inglês Instrumental para cursos técnicos de Eletrotécnica e cursos de graduação em Engenharia Elétrica. A Eletrotécnica é uma área da Engenharia Elétrica que lida com instalações de redes elétricas, transformadores, circuitos elétricos residenciais, industriais, etc. A principal necessidade desse público-alvo para a realização de seu trabalho é a leitura de documentos técnicos chamados *datasheets*, que contêm especificações sobre o funcionamento de dispositivos elétricos (Dewey, 1998). Esses documentos se encontram em sua maioria escritos em Língua Inglesa. A pesquisa parte da premissa de que o protótipo de glossário deve ser desenvolvido a partir de preceitos de uma metodologia que chamamos Terminografia Didático-Pedagógica, prática terminográfica baseada na triangulação entre preceitos da Teoria da Terminologia de perspectiva Textual (Ciaspucio, 1998, 2003; Finatto, 2004; Krieger, 2004; Bourigault e Slodzian, 2004); da Teoria Sócio-Cognitiva da Terminologia (Temmermann, 1997, 2000, 2004); e da Linguística de Corpus (Biber, 1988; Sinclair, 1991; Berber Sardinha, 2004). A Terminografia Didático-Pedagógica (TD-P) combina dados extraídos de gêneros textuais (Swales, 1990; Bazerman, 2005) relevantes ao ambiente de ensino com dados obtidos em coleta de dados diretamente com os participantes desse ambiente. O usuário direto dessa metodologia deve ser o professor/pesquisador de ESP e/ou o pesquisador de Terminografia/Terminologia, pois a utilizará para produzir materiais e ferramentas que o auxiliem em seu trabalho em sala de aula. O usuário indireto dessa metodologia será o aprendiz da área técnica com diferentes níveis de proficiência em Língua Inglesa. Metodologia: Passo 1: observar e detectar as dificuldades com o vocabulário e em relação à leitura dos *datasheets* no ambiente de ensino; Passo 2: coletar dados do gênero textual escolhido, no caso os *datasheets*; Passo 3: coletar dados com os alunos; Passo 4: comparação dos dados de acordo com critérios estabelecidos na observação das dificuldades; Passo 5: montagem do protótipo de acordo com dados extraídos da comparação. Através dos resultados, concluímos que a combinação de dados proposta pela metodologia da Terminografia Didático-Pedagógica pode auxiliar na montagem de uma ferramenta mais direcionada ao público-alvo que a utilizará algum dia, justamente porque resultou em dados diversos do que uma pesquisa exclusivamente com ferramentas que realizam a seleção de candidatos a termos poderia fornecer.

34

PORTUGUÊS PARA FINS ACADÊMICOS EM UMA ABORDAGEM BASEADA EM *CORPUS*: UM ESTUDO SOBRE VERBOS

Lucas Zambrano Rollsing, Cristina Becker Lopes Perna e Lucelene Lopes
Escola de Humanidades – PUCRS; UNDL Foundation – Geneva
lucas.rollsing@acad.pucrs.br, cperna@pucrs.br, l.lopes@undlfoundation.org

Abstract: *The study hereby intends to present an interface thesis between Corpus Linguistics and Natural Language Processing (NLP) in progress at PUCRS. This project consists in the exploration, through a written corpus, of theses and dissertations from the PUCRS Graduate Program in Linguistics, through a software called ExATO (Lopes, 2012). From this tool, we can have a series of linguistic resources that allow the continuation of our analysis, such as extracted verb lists, concordance and verb clouds, among others. We intend to explore the verbs used within the academic sphere, in order to base a proposal of teaching of Portuguese as an Additional Language, with a view to enhance the proficiency within the academic genre by non-Portuguese speakers.*

Resumo: O presente estudo é uma dissertação em interface entre Linguística de Corpus e Processamento de Linguagem Natural em andamento na PUCRS. Via *corpus* escrito, teses e dissertações em Linguística oriundas do PPGL da referida instituição foram coletadas e analisadas através do software chamado ExATO (Lopes, 2012). A partir dessa ferramenta podemos analisar listas de verbos extraídos, concordanciador e nuvens de verbos, dentre outros recursos. A fim de embasarmos uma proposta de ensino de Português como Língua Adicional, com vistas à proficiência dentro do gênero acadêmico por alunos não falantes de LP, nosso objeto de análise, portanto, torna-se o registro acadêmico e os verbos nele utilizados.

35

CONSTRUÇÕES DE TÓPICO NO PORTUGUÊS FALADO EM ÁREAS INDÍGENAS

Wlianna Silva de Araújo e Edivalda Alves Araújo
Universidade Federal da Bahia (UFBA)
wliannasaraujo@outlook.com; edivalda.araujo@gmail.com

Esse trabalho tem como objetivo identificar a realização de construções de tópico em dados orais do português falado em áreas indígenas brasileiras, com a intenção de levantar a frequência dessas construções e suas características sintáticas, trazendo mais contribuições para os trabalhos nessa área. A análise aqui apresentada tem aporte teórico em Pontes (1987) e Omitido (2006), em relação às construções de tópico, e outros autores, como: Decat (1989), Galves (2001) e Omitido (2009a, 2009b), em relação ao português brasileiro.

O OBJETO DIRETO EM CONSTRUÇÕES DE TÓPICO DO SÉCULO XIX

Cristovão Mascarenhas Cordeiro e Edivalda Alves Araújo
Universidade Federal da Bahia (UFBA)
chris_mascarenhas@live.com; edivalda.araujo@gmail.com

Muitos pesquisadores têm se dedicado a estudos voltados para as construções de tópico no português brasileiro (PB). Entre esses linguistas, destaca-se Eunice Pontes (1987), cujos trabalhos têm desencadeado discussões acerca da construção sintática dessa língua, a partir da hipótese de que o PB estaria se definindo como uma língua com proeminência de tópico (elementos deslocados à esquerda da oração). Segundo Pontes (1987), quando se começa a observar a língua coloquial espontânea no uso diário, fica-se surpreso com a quantidade de vezes em que ocorrem construções com tópico. Com este trabalho, buscamos rastrear diacronicamente as construções de tópico no PB para identificar se são produtivas a ponto de caracterizá-lo como uma língua de proeminência de tópico já em períodos pretéritos (século XIX) e quais os fatos sintáticos que podem ser identificados nessas construções. A partir da problematização proposta, levanta-se a hipótese de que, para se desenvolver como língua com proeminência de tópico, o PB já apresentava essa tendência em textos produzidos em séculos anteriores, como o XIX, por exemplo. Assim, esta pesquisa objetiva analisar construções sintáticas do PB dos séculos XIX a XXI, envolvendo o deslocamento do objeto direto para a posição de tópico, para identificar as características e os contextos de ocorrência dessas construções. É importante que o conceito de tópico seja minimamente esclarecido para o entendimento de nossa pesquisa. De forma breve, podemos dizer que a definição de tópico pode ser estabelecida a partir de dois grandes grupos de características: sintáticas e semânticas. No que tange a sintaxe, está relacionada à ordem em que as palavras compõem o discurso: é o elemento deslocado à esquerda na oração. A sintaxe é responsável por oferecer o local para o tópico acontecer; a semântica, por sua vez, restringe o SN, não é qualquer SN que pode ser considerado tópico, mas sim aqueles que possuem os traços: (+ definido); (+ específico); (+ identificável); (+ ativo) (cf. Araújo, 2006). Dessa forma, o tópico (elemento deslocado à esquerda), sob a perspectiva semântica, possui características como: referencialidade, identificabilidade, definitude, etc. Para a primeira parte desse estudo, optou-se por observar obras do século XIX. Entre essas, foram analisadas algumas peças de teatro: "O noviço", Martins Pena (1845), "As desgraças de uma criança", Martins Pena (1846); "Caiu o mistério", França Júnior (1883); os romances "O Cortiço" (1890), de Aluísio de Azevedo; "Triste fim de Policarpo Quaresma" (1915), de Lima Barreto; "Quincas Borba" (1881) e "Dom Casmurro" (1899), de Machado de Assis (1891); três contos desse mesmo autor: "Teoria do medalhão" (1881); "A Cartomante" (1884); "Missa do Galo" (1893). Escolhemos o século XIX pois esse período reflete o caráter intermediário do fenômeno relacionado ao ordenamento dos constituintes na sentença. Segundo Tarallo (1996), é a partir do século XIX que o processo de formalização do fenômeno se intensifica e a mudança em relação ao "enrijecimento" da ordem SVO se concretiza; é ainda nesse período que uma gramática do português brasileiro começa a ser delineada, uma gramática independente da gramática de Portugal. O século XIX é, portanto, o marco inicial de suma relevância para o nosso trabalho. A partir da análise dos dados, objetivou-se verificar as construções exis-

tentes e as diferenças ou semelhanças entre elas para, hipoteticamente, capturar as características sintáticas das construções de tópico envolvendo a posição de objeto direto nesse período, bem como avaliar a classificação do português brasileiro como língua direcionada para o discurso. Curiosamente, os autores analisados não apresentaram em seus textos construções de tópico como as apresentadas por Pontes (1987) ou envolvendo o objeto direto, à exceção de Martins Pena. Os textos deste autor são considerados de cunho mais popular e de linguagem espontânea, contexto considerado favorável para as construções de tópico. Na verdade, a análise dos seus textos demonstra que o autor era formal em suas construções. Talvez a sua irreverência estivesse na abordagem de alguns assuntos ou de temas, considerados tabus ou polêmicos para a época. As suas construções sintáticas, ao contrário, não demonstram irreverência, mas antes conservadorismo. Diferentemente dos outros autores analisados, que não apresentaram construções de tópico, Martins Pena destaca-se por trazer construção de tópico envolvendo o objeto direto: a do tipo topicalização V2, como no exemplo: "Paciência, preciso eu também ter, e muita" (MP. O Noviço, 1845). Esse tipo de construção, segundo Ribeiro (1996), apresenta como característica principal a inversão na ordem dos constituintes: o verbo ocupa a segunda posição e há uma inversão verbo-sujeito. São construções típicas do português arcaico, mas, embora com pouca frequência, foram encontrados resquícios dessas construções no português brasileiro que perduraram até o início do século XX (cf. Ribeiro, 1996). Segundo Ribeiro (1996), o sujeito pós-verbal pode ser resultante do movimento de verbo para Cº. Seria um tipo de frontalização do verbo com deslocamento de algum outro constituinte para a posição inicial da sentença, imediatamente antes do verbo, originando a ordem XV ou V2, caracterizando na linguística o que se denomina de "inversão germânica". Encontramos esse tipo de inversão nos dados de Martins Pena, o que revela uma singularidade em sua sintaxe, se comparada a de outros escritores do mesmo período. O estudo tem nos mostrado que as construções de tópico sofrem evolução ao longo da história da língua. São dinâmicas. Ou seja, construções encontradas no século XIX não são encontradas no século XXI e vice-versa. Supomos que as construções de tópico estão de acordo com o percurso sintático da língua. As características sintáticas do português brasileiro eram diferentes no século XIX: havia mais morfologia, mais sujeito nulo, mais movimento de verbo; consequentemente, era possível encontrar construções de tópico do tipo V2. Com a perda da morfologia, há mudanças sintáticas no PB que envolvem também as construções de tópico, mudando as suas ocorrências. O desenvolvimento deste trabalho tem permitido identificar a evolução dessas construções no português brasileiro. Espera-se, assim, que os resultados dessa pesquisa contribuam para ampliar as discussões relacionadas aos estudos diacrônicos da sintaxe dessa língua.

49

BRWAC: UM GRANDE *CORPUS* PARA GRANDES DESAFIOS

Jorge Alberto Wagner Filho, Rodrigo Wilkens e Aline Villavicencio
Universidade Federal do Rio Grande do Sul (UFRGS)
jawfilho@inf.ufrgs.br; rodrigo.wilkens@inf.ufrgs.br; avillavicencio@inf.ufrgs.br

Corpora de grande volume tornam-se cada vez mais necessários em diferentes tarefas de Processamento de Linguagem Natural. Isto levou pesquisadores a buscarem

em uma das maiores fontes de conteúdo dos tempos atuais, a *web*, materiais para complementar seus repositórios tradicionais. Neste trabalho, apresentamos a construção de um *corpus web* de grandes proporções para a variante brasileira da língua portuguesa, visando atingir um tamanho comparável ao estado da arte de outros idiomas. O corpus obtido é composto por 2,7 bilhões de palavras, e foi completamente anotado com um analisador sintático.

50

PLURISIGNIFICAÇÃO EM TRANSITIVIDADE

Rodrigo Esteves de Lima Lopes
Universidade Estadual de Campinas (UNICAMP)
rll307@unicamp.br

Este artigo tem por objetivo discutir critérios para classificação de processos dentro do escopo da Gramática Sistêmico-Funcional, a partir de uma abordagem baseada em *corpus*. Argumenta-se que as características semânticas dos processos podem ser mapeadas a partir de suas colocações, estabelecendo, assim, padrões gramaticais que poderiam servir como parâmetro de análise.

62

O USO DE *CORPORA ONLINE* MULTIMODAIS PARA ENCONTRAR A EQUIVALÊNCIA DE TERMOS EM INGLÊS TENDO COMO BASE O DICIONÁRIO OLÍMPICO

Caroline Cougo
Universidade do Vale do Rio dos Sinos (Unisinos)
caroline.cougo@gmail.com

O presente trabalho visa à análise do trabalho de *corpora* online realizado para buscar equivalentes em inglês para termos do português durante a confecção do Dicionário Olímpico, o qual foi realizado na Unisinos pelo grupo Semantec e coordenado pela Profa. Dra. Rove Chishman. Para começar, é importante introduzir e explicar o funcionamento do dicionário, que foi inspirado nos jogos olímpicos de 2016, tendo como base referencial a Semântica de *Frames*, de Charles Fillmore. O autor descreve um *frame* como "qualquer sistema de conceitos relacionados de tal maneira que para entender qualquer um deles é preciso entender a estrutura que os comporta como um todo" (Fillmore, 1976). Juntamente com essa teoria, trabalha-se com a Lexicografia Computacional e a Linguística de Corpus, que é "uma coletânea de textos naturais (*naturally occuring*), escolhidos para caracterizar o estado ou variedade da linguagem" (Sinclair apud Sardinha, 2000, p. 16). Esta foi usada para coletar equivalentes em inglês e exemplos de uso, e, para isso, usamos programas como AntConc e *Sketch Engine*. Para fazer uma coleta de unidades lexicais e exemplos de uso nesses programas, tínhamos como primeiro objetivo fazer a coleta de materiais escritos de cada esporte. Primeiramente, deveríamos selecionar *links* que fossem considerados fontes de dados; a saber, *sites* sobre esportes, *sites* de federações, periódicos especializados e

manuais. Nós tínhamos dois tipos de *corpora*: o *corpus* de consulta e o *corpus* de pesquisa. O primeiro era um documento do Word com uma tabela de *links* que julgávamos como os mais informativos, os quais consultávamos cada vez que tínhamos alguma dúvida sobre o esporte pesquisado. O segundo, *corpus* de pesquisa, era coletado também no Word, porém era um texto corrido, isto é, continha apenas os textos, sem data de postagem, título ou qualquer outra informação que por ventura os circundassem. Assim deveria ser porque esse *corpus* seria formatado para txt e seria rodado no programa AntConc. Um fato relevante em relação aos *corpora* de consulta e pesquisa é que o *corpus* de consulta permitia qualquer tipo de texto que trouxesse informações que o pesquisador achasse relevante, enquanto o de pesquisa permitia apenas textos do tipo *match reports* – ou seja, notícias que descrevessem as partidas, com o tempo verbal geralmente no passado, e contendo termos de cada esporte. Por exemplo, um *corpus* de pesquisa relevante do futebol, teria uma narrativa do tipo: "O jogador fez uma bicicleta e chutou em direção ao gol". Os termos "bicicleta" e "gol" seriam termos específicos do futebol e o tempo verbal no passado. Os esportes mais populares, como futebol, vôlei ou handebol deveriam ter, no mínimo 500.000 caracteres, e os menos populares, como esgrima, 150.000. A dificuldade que encontramos, entretanto, foi que os esportes menos populares não apresentavam *match reports* suficientes para preencher o mínimo do *corpus* de pesquisa. O mínimo de 500.000 caracteres para esportes mais populares, entretanto, eram facilmente preenchidos. No meio da produção, então, o grupo de pesquisa e o grupo de tradução desistiram dessa coleta de *corpus* por meio de programas e começaram a coletar unidades lexicais, tradução e exemplo de uso a partir de *corpora* online multimodais, como Google News e Imagens, narrativas de jogos no Youtube e até mesmo o Linguee. A lista de palavras e de coletas de exemplo em inglês não passava de 200 em cada esporte, e, para cada um deles, foi organizada uma planilha no Excel com as unidades lexicais em português, contendo, ao lado, seu equivalente em inglês. Caso não houvesse um equivalente em inglês para uma palavra, uma nota ao lado explicava seu significado. Foram encontradas equivalentes em inglês para a grande maioria da lista de ULs. O trabalho de pesquisa do grupo de tradução a partir de *corpora* na *web* foi realizado por um grupo de quatro pessoas e o tempo médio de término de uma planilha com equivalentes e exemplos de uso girava em torno de cinco a seis horas, sendo cada pessoa responsável por um esporte de cada vez. A pesquisa era feita no modelo termo a traduzir + nome do esporte em inglês. Quando havia um termo com mais de uma palavra, pesquisava-se no Google usando aspas entre ele, pois sabíamos que, assim, a ferramenta de pesquisa online teria como resultado o termo entre aspas e não cada palavra individualmente. Afinal, essa pesquisa online foi efetiva, pois a grande maioria das unidades lexicais pesquisadas foram encontrados com exemplos reais de uso. A partir do trabalho realizado, percebe-se que o *web corpora* pode ser muito útil e eficiente para a pesquisa de equivalentes de palavra e que a teoria da semântica de *frames* é válida e tem muito a contribuir para a construção de um dicionário online. Conclui-se que houve uma limitação do uso do AntConc para os propósitos que tínhamos em mente, visto que alguns esportes olímpicos não contêm muito material textual na internet e precisávamos de uma boa base para a coleta. Os *corpora* multimodais foram de extrema ajuda, considerando que, se não houvesse um equivalente na pesquisa do Google, por vezes conseguimos achá-lo em vídeos e narrativas do esporte em questão. A combinação da semântica de *frames*, *web corpus* e lexicografia permitiu que um dicionário completo e que muito tem a contribuir para o conhecimento sobre esportes e linguística fosse produzido.

UM ESTUDO SOBRE OS PARTICÍPIOS COM BASE EM *CORPUS*

Luiza Trugo e Cláudia Freitas
Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO)
claudiafreitas@puc-rio.br

Relatamos aqui os resultados de um estudo sobre as formas participiais do português, motivado pela tarefa de (re)anotação do *corpus* Mac-Morpho (Aluísio et al., 2003) segundo o *tagset* do projeto Universal Dependencies (http://universaldependencies.org/). A motivação para a (re)anotação do MacMorpho, por sua vez, vem do desejo de realizar estudos linguísticos que tematizem o impacto de diferentes *tagsets* (de pos, mas não só) em sistemas de PLN. Considerando o português contemporâneo, temos 2 grandes *corpora* de referência, isto é, anotados e revistos: o Bosque (Afonso et al., 2002; Freitas et al., 2008) e o Mac-Morpho (Aluísio et al., 2003; Fonseca et al., 2003; Fonseca et al., 2015). O Bosque, parte do projeto Floresta Sintá(c)tica e o primeiro treebank do português, além da anotação de pos, também contém anotação sintática, ambas revistas por linguistas. Porém, com cerca de 200 mil palavras, das variantes do Brasil e de Portugal, é considerado um *corpus* relativamente pequeno. O MacMorpho contém apenas anotação de pos, mas por outro lado é de dimensões maiores, dado relevante quando se pensa no treino de sistemas em abordagens de aprendizado de máquina. O *tagset* de pos do projeto Universal Dependencies (UD) vem da proposta do Google de *tags* universais (Petrov et al., 2012). Atualmente, já integram o projeto mais de 58 *corpora* anotados para mais de 40 línguas, português incluído. No contexto do UD, o material para a língua portuguesa refere-se ao Bosque: em Petrov et al. (2012) relata-se o processo de conversão automática do Bosque para o *tagset* UD, e está em andamento a disponibilização de uma versão do mesmo material, mas com uma conversão e revisão feitas de maneira semiautomática, levando em conta uma série de análises e resoluções que só poderiam ser feitas com uma análise linguística cuidadosa. Ou seja, já temos o Bosque com dois *tagsets*: *tagset* UD e *tagset* do Bosque. Gostaríamos que o MacMorpho, pelo seu tamanho, qualidade e farta utilização no aprendizado de pos de língua portuguesa, também pudesse dispor de mais de um *tagset*, e por isso a tarefa de (re)anotação. O *tagset* do MacMorpho é bem mais granular que o do projeto UD. Algumas classes se alinham de maneira imediata, mas isso não funciona para todas, como é o caso das formas participiais. Já sabendo da ambiguidade que assumem, podendo ser classificadas ora como verbos e ora como adjetivos, a equipe de anotação do Mac-Morpho optou por manter os particípios como uma classe a parte, chamada PCP. Na conversão para o *tagset* UD, é preciso distribuir os PCPs nas classes ADJ, VERB e também NOUN e ADV. O grande desafio, no entanto, é distinguir entre ADJ e VERB, fato já atestado em uma série de gramáticas, e mencionado de forma explícita por Villalva e Silvestre (2014, p. 164, grifo nosso): "Problemas suscitados pela classificação de palavras como o infinitivo, o gerúndio e o particípio [...] indicam que as classes de palavras tradicionalmente reconhecidas constituem matéria que pode e deve ainda vir a ser discutida". Ao longo do processo de conversão, a inexistência de um tratamento consistente e de uma posição consensual em gramáticas, teorias linguísticas e na própria documentação UD (que afirma apenas:

"Note that participles are word forms that may share properties and usage of adjectives and verbs. Depending on language and context, they may be classified as either VERB or ADJ") nos obrigou, em muitos casos, a um tratamento individual das formas participiais, que não foi feito sem hesitação em uma série de casos. A fim de validar nossas opções linguísticas, conduzimos um estudo sobre a classificação das formas participiais tendo como foco profissionais (professores e pesquisadores) da área de Letras com experiência em gramática do português. O estudo foi feito utilizando a ferramenta Rêve (http:www.linguateca.pt/Reve), e as frases, bem como as análises (anotação), estão públicas na página da ferramenta. No estudo, selecionamos frases que continham formas participiais que consideramos de difícil classificação; formas participiais que consideramos mais claramente adjetivais e formas participiais que consideramos mais claramente verbais. Pedimos aos 10 participantes que classificassem as formas em questão como V, ADJ, NÃO SEI, OUTRO, sempre levando em conta o contexto e também a possibilidade de classificações múltiplas. Os resultados foram na direção que esperávamos, com uma imensa divergência na classificação das formas que consideramos difíceis – cf. tabelas 1, 2 e 3. Ou seja, diferentemente da tarefa de anotação clássica, em que se busca verificar a concordância na anotação, o que queríamos aqui era justamente verificar a discordância nas análises, pois nossa sensação era estarmos diante de uma classe nada homogênea. Estamos habituados a pensar nas classes de palavras como categorias discretas, e a anotação linguística reforça essa visão, mas os resultados sugerem que, ao menos em certos casos, essa pode não ser a melhor solução. De fato, um comportamento não discreto, não categórico, das classes de palavras é também um ponto relevante abordado em Manning (2015). E, de um ponto de vista da historiografia linguística, é de grande relevância lembrar que nem sempre os particípios foram uma forma verbal. Desde os primeiros filósofos estoicos (301 a.C.), muitos estudiosos consideravam particípios "nomes verbais", "verbos com casos", "verbos participiais", dentre outros, o que enfatiza suas características mais verbais — apesar de os particípios também se flexionarem em caso, gênero e número, como os nomes. Vale lembrar que, quando foi feita a tradução das classes de palavras do grego para o latim, o particípio (participium) recebeu esse nome por "participar" tanto da classe dos nomes quanto da dos verbos. Considerava-se que particípios "participavam" da classe dos nomes devido à ausência de Modo e à presença de Caso e Gênero, e da classe dos verbos devido ao fato de apresentarem Tempo e flexões derivadas do verbo (Aurux, 2002). Por fim, a pesquisa em *corpus* trouxe informações para lidar com o fenômeno das formas participiais que até então não estavam disponíveis, enriquecendo também a descrição gramatical. A literatura linguística sobre o tema está baseada em intuições e exemplos controlados, sem levar em consideração (por falta de meios para fazê-lo) como têm sido usadas as formas participiais: em que contextos costumam aparecer, como seus papéis mudam de acordo com os elementos que as cercam, como variam de acordo com o gênero e o registro linguísticos, dentre outras informações relevantes para uma abordagem mais rica e abrangente do fenômeno. E um olhar linguístico-descritivo aguçado na tarefa de anotação para PLN mostra o quanto este último pode contribuir para ampliar nossa visão da língua.

MODELAGEM DE CONSTRUÇÕES DE ESTRUTURA ARGUMENTAL NO *CONSTRUCTION* DA FRAMENET BRASIL: DESDOBRAMENTOS PARA O PROCESSAMENTO DE LÍNGUA NATURAL

Vânia de Almeida e Tiago Torrent
Universidade Federal de Juiz de Fora (UFJF)
vaniaga87@hotmail.com, tiagotorrent@gmail.com

O presente trabalho se enquadra nas discussões teórico-metodológicas da FrameNet Brasil (FN-Br), a partir de uma abordagem interdisciplinar entre Linguística e Ciência da Computação, a qual resulta na construção de recursos lexicais e dicionários eletrônicos e na identificação e catalogação de construções do PB para o Processamento de Língua Natural. A FN-Br, seguindo o que é feito na FrameNet (FN), sediada no International Computer Sciences Institute, em Berkeley (http://framenet.icsi.berkeley.edu), em relação ao Inglês, ocupa-se do desenvolvimento de dois grandes recursos computacionais: um Lexicon e um Constructicon. O Lexicon tem como objetivos (a) descrever Unidades Lexicais (ULs) a partir dos *frames* que evocam e identificá-los em termos dos Elementos de *Frame* (EFs) que estão expressos em cada sintagma; (b) disponibilizar os resultados da anotação na forma de entradas lexicais, as quais sumarizam padrões de valência de cada UL. O desenvolvimento do Constructicon na FrameNet Brasil assenta-se sobre o objetivo principal de implementar um recurso que, em complemento ao Lexicon, possa explicar os fenômenos que extrapolam as possibilidades semânticas e sintáticas de unidades lexicais e, assim, descrever construções em termos de suas propriedades gramaticais e seu potencial semântico, de modo compatível com os pressupostos teóricos da Berkeley Construction Grammar (Kay e Fillmore, 1988; Fillmore, 2008; 2013), incorporando descrições interpretáveis computacionalmente para cada construção. Diante dessa gama de recursos, este trabalho tem por objetivos fornecer a modelagem linguístico-computacional necessária para o tratamento das construções Transitiva Direta Ativa, Ergativa e de Argumento Cindido no Constructicon do PB e apresentar uma tarefa de descoberta de construções por máquina cujo objetivo é o de identificar instâncias das construções em questão. A construção Transitiva Direta Ativa caracterizada formalmente pelo esquema [SN [V SN]] apresenta uma estrutura biargumental em que os dois sintagmas nominais, representados pelo sujeito e objeto direto, indicam quem praticou uma ação e quem sofreu a ação, respectivamente, como em João rasgou o papel. Já a construção Ergativa apresenta uma estrutura monoargumental [SN [V]] em que o sujeito é paciente, ou seja, não pratica nenhuma ação, mas, sim, sofre o efeito do evento expresso pelo verbo, como encontrado em O vaso quebrou. As construções de Argumento Cindido – CACs – (Sampaio, 2010) são construções sintaticamente semelhantes às transitivas, com estrutura biargumental, porém semanticamente semelhantes às ergartivas, indicando um evento sofrido por uma entidade, como em O carro furou o pneu. Ainda sobre a CAC, ressalta-se a existência de uma relação Todo-Parte entre o sujeito e o objeto direto, respectivamente. Para modelar adequadamente as características sintático-semânticas dessas três construções foi necessário considerar (a) quais os verbos mais prototípicos encontrados nessas construções; (b) os *frames* que eles evocam e a qual

relação entre os elementos desses *frames* e os argumentos do verbo; (c) o *frame* mais genérico evocado pelas construções e a relação entre os elementos desse *frame* e os argumentos das construções; (d) as relações de herança entre as construções. A nossa proposta se insere dentro da fase representacional no âmbito da Linguística Computacional, conforme a metodologia proposta por Dias da Silva (1996), uma vez que modela computacionalmente as construções Transitiva Direta Ativa, Ergativa e CAC no Constructicon e, posteriormente, apresenta e discute os resultados de uma tarefa de descoberta construcional por máquina através da técnica de travessia em redes chamada Ativação Propagada (AP) ou Spreading Activation. A modelagem construcional permite apresentarmos de que forma as propriedades das construções Transitiva Direta Ativa, Ergativa e CAC estão relacionadas com outras construções que contribuem para sua definição, seja através de herança, ou composição. Com a elaboração de uma rede, mapeamos como se constituem as relações de herança entre construções e percebemos que as construções em estudo são todas herdeiras de uma construção mais genérica Sujeito_Predicado, que, por sua vez, é herdeira da Construção X_Núcleo. A distinção entre os Elementos da Construção (ECs) das construções Transitiva Direta Ativa, CAC e Ergativa foi corresponsável pela modelagem dessas construções no Constructicon, uma vez que foi possível unificar a contraparte semântica e contraparte sintática das construções através do mapeamento dos ECs para os EFs dos *frames* evocados por elas. Assim, temos a CAC e a Ergativa semanticamente sinônimas, porém com a CAC apresentando uma relação parte-todo entre seus argumentos o que não é observado na Transitiva Direta Ativa que é sintaticamente semelhante a ela. Na fase implementacional, que foi adaptada de Matos (2014), discutimos os resultados de uma tarefa de descoberta construcional por máquina através do *parser* em desenvolvimento pela FN-Br, o Constructional Analyzer using Relations among Multiple AVMs (Carma). Para essa discussão foi realizado, primeiramente, um teste de julgamento semântico com o objetivo de estabelecimento de um *gold standard* para verificar se as construções em estudo possuem um padrão semântico que pode ser facilmente identificado por falantes do PB, e fornecer evidência quantitativa para sustentar essa hipótese. O teste de similaridade semântica foi apresentado por meio de uma pesquisa virtual, criada e aplicada através da ferramenta *web* Survey Monkey. O teste continha 40 sentenças, sendo 10 licenciadas pela construção Transitiva Direta Ativa, 10 pela Ergativa, 10 pela CAC e 10 sentenças distratoras. A partir da realização desse teste para o estabelecimento de um *gold standard*, as mesmas sentenças das construções Transitiva Direta Ativa, Ergativa, e CAC foram submetidas ao Carma. O teste de similaridade semântica indicou o reconhecimento, por parte dos falantes do PB, dos padrões das construções em estudo. A partir da comparação dos resultados, constatamos que o *parser* reconheceu os padrões construcionais da mesma maneira que os falantes nativos do PB. Os resultados dessa tarefa apontam que a representação realizada no Constructicon para as construções permitiu que o sistema identificasse, satisfatoriamente, 29 das 30 sentenças utilizadas no experimento, aproximando-se do *gold standard* definido a partir do julgamento semântico das mesmas 30 sentenças por falantes nativos do PB.

AOS PROFESSORES, AS COLOCAÇÕES

Andréa Geroldo dos Santos
Universidade de São Paulo
andrea.geroldo@gmail.com

Introdução: Este trabalho, um recorte de nossa pesquisa de Doutorado, tem como objetivo demonstrar como temos desenvolvido oficinas para professores de Inglês do Ensino Fundamental II e Ensino Médio, visando a sensibilizá-los quanto à importância de se ensinar colocações e à utilidade das ferramentas baseadas em Linguística de Corpus para realizar tal tarefa. Para isso, valemo-nos do gênero discursivo "livro didático", já que esse tipo de material se constitui em fonte de conhecimento no processo de ensino-aprendizagem no Brasil e, muitas vezes, é a única fonte de leitura para os professores (Coracini, 1999; Souza, 2002). Além disso, desenvolvemos essas oficinas como parte de nosso trabalho como editora de livros didáticos para o ensino de inglês para brasileiros, no Sistema Mackenzie de Ensino (doravante SME), patrocinado pela Universidade Presbiteriana Mackenzie. Esse sistema de ensino elabora o material didático não só para os colégios Mackenzie, mas também para mais de 200 escolas no Brasil. O livro didático de Inglês é constituído de unidades temáticas que apresentam textos autênticos adaptados (em geral, textos jornalísticos) e informado por *corpus*, valendo-se, principalmente, de ferramentas disponíveis no COCA: KWIC, Key Word in Context – para a elaboração de exercícios lexicais e gramaticais; e Word and Phrase – para a análise dos textos autênticos utilizados, no que se refere às colocações e palavras mais frequentes. Contudo, tal abordagem não é a que se encontra normalmente no mercado editorial de ELT (English Language Teaching) em geral (Burton, 2012; Tomlinson, 2003), muito menos no Brasil, onde os livros nacionais ainda tendem a favorecer textos pedagógicos curtos (criados pelos autores), listas de palavras e apresentação de regras. Tal fato, aliado ao desafio da pesquisa e adaptação constantes quando se trabalha com textos autênticos, parecem explicar por que muitos professores acabam por rejeitar materiais assim elaborados (Hanna, 2012). Outro fator também seria a formação muitas vezes insuficiente dos docentes. Naturalmente, esses fatores têm influenciado o *feedback* que recebemos em relação ao material do SME. Por essa razão, elaboramos oficinas de curta duração a fim de: a) verificar o modo como os docentes abordam o texto autêntico; b) chamar a atenção quanto à importância de se trabalhar a convencionalidade, em especial, as colocações; c) introduzir ou revisar os elementos básicos da Linguística de Corpus; d) propor que os docentes elaborem material didático com vistas ao uso de textos autênticos e ferramentas baseadas em *corpus*. Dessas oficinas, realizamos dois pilotos relacionados aos itens "a" e "b" supramencionados, os quais descreveremos neste trabalho.

Metodologia: Para as oficinas, trabalhamos com dois grupos diferentes, A (cinco docentes) e B (sete docentes), todos do Estado de São Paulo e profissionais das escolas que utilizam o material do SME. Em diferentes datas, os dois grupos trabalharam com dois textos autênticos adaptados: grupo A, com o texto Myths and legends: the Loch Ness monster; grupo B, com o texto The impact of nutrition on your health. Esses textos fazem parte do livro do 8º Ano do SME que estava em elaboração e, portanto, com o qual

os docentes ainda não tinham tido contato. Feita essa divisão, as oficinas ocorreram da seguinte forma: Oficina I (doravante I): os grupos A e B leram os textos e propuseram dois ou três exercícios que explorassem a compreensão linguística (lexical) de seus alunos. Oficina II (doravante II): A e B observaram como os textos em que eles trabalharam em I foram efetivamente abordados no livro do 8º Ano – ou seja, quais exercícios foram propostos, levando-se em consideração a convencionalidade (mais especificamente, as colocações presentes nos textos). Ainda em II, os docentes tiveram o primeiro contato com ferramentas da Linguística de Corpus, já que os exercícios propostos para a compreensão linguística dos textos autênticos do livro de Inglês do 8º Ano são: a) baseados na análise obtida na ferramenta Word and Phrase (doravante WP), no tocante à frequência das palavras e às colocações; b) elaborados levando-se em consideração o seguinte: o ensino DDL (Data-Driven Learning), ou Aprendizagem Direcionada por Dados, que propõe que o aluno infira regras e usos de fenômenos linguísticos a partir da análise de corpus (Johns, 1991); o uso de linhas de concordância para o ensino (Tribble e Jones 1997; Berber Sardinha, 2004; Gavioli, 2005;); os "três I's" – Illustration, Interaction and Induction (Carter e McCarthy, cf. Xiao e McEnery, 2005); a Modelagem, que postula modelar os dados dos padrões a serem ensinados, com base em dados autênticos – ou seja, *corpora* (Carter, 1998). Todavia, considerando-se a quantidade de informações e a duração de II (quatro horas), apenas listamos a base teórica da abordagem do material – e que será aprofundada nas próximas oficinas, focando também na ferramenta WP.

Resultados: Em I, tanto os grupos A quanto B propuseram exercícios em que os alunos deveriam sublinhar ou copiar as palavras cognatas, por exemplo: Texto A: *famous*, *enormous* e *creature*. Texto B: *important*, *reduce* e *adults*. Além do exercício de reconhecimento dos cognatos, A ainda solicitou que os alunos lessem o que haviam sublinhado e apontassem a ideia principal do texto. B, todavia, não deu sequência ao trabalho com os cognatos, preferindo trabalhar com outra lista de palavras do texto: *food*, *today*, *tomorrow*, *future*, *health*(*y*), *dietary*, *child*/*adulthood*, *carry* e *benefits*. Embora o levantamento dos cognatos seja uma técnica importante para auxiliar o aluno a compreender um texto, pudemos notar que a preocupação maior dos docentes era trabalhar "listas" – ou seja, reconhecer palavras isoladas, não explorando suas (possíveis) relações, tampouco sua classe morfológica e/ou gramatical. Tal preocupação ficou evidente principalmente no caso do grupo B, que abandonou a primeira lista (cognatos) por uma segunda. Quando perguntados qual o critério utilizado para a escolha dessa segunda lista, os docentes informaram que julgavam tais palavras "importantes", sem elaborar o porquê. Insistimos que *food*, *today* e *tomorrow* são bem conhecidas, já que os alunos do 8º Ano não são iniciantes em Inglês. Não obtivemos resposta à nossa indagação. As outras questões propostas pelos dois grupos saíram do escopo que havíamos delineado (exercícios de compreensão linguística, com foco no léxico): A solicitou que os alunos sublinhassem todos os verbos no Simple Past, classificando-os em regular e irregular. Já B apresentou quatro perguntas relacionadas à compreensão das ideias do texto. Em II, distribuímos os textos novamente, e sua versão já analisada na ferramenta WP. Os docentes de ambos os grupos ficaram "maravilhados" quando viram os textos com as palavras destacadas, de acordo com a frequência. Disseram ainda que tal organização facilitava a visualização de expressões, algo que eles não haviam notado quando do trabalho com os textos em I. Por exemplo: Em A: *spread rumours*, *take a photograph*, *conduct a survey* e *conclusive proof*. Em B: (*un*)*healthy lifestyle*/*eating ha-*

bits/weight e *chronic/heart disease*(*s*). Apresentamos, então, como essas colocações foram abordadas no livro do SME, demonstrando o modo como usamos o COCA para nos auxiliar na busca por outras possíveis colocações e outros exemplos. Finalmente, os docentes tiveram acesso a alguns dos exercícios, já diagramados no livro. Os docentes afirmaram, então, que passaram a entender qual era a proposta do livro do SME: abordar poucas palavras, em vez de listas, mas aprofundar seu estudo, por meio da análise do modo como elas se relacionam com as outras palavras à sua volta. E em um contexto autêntico de uso. Para a oficina seguinte (III), pedimos que os docentes escolhessem um texto autêntico curto, fizessem sua análise no WP e elaborassem duas ou três questões para analisar a compreensão linguística desse texto. Esperamos que tais questões possam refletir o que os docentes aprenderam no II. Sabemos que esse é apenas um recorte do trabalho que teremos de fazer com os professores das mais de duzentas escolas que utilizam esse material constituído de textos autênticos e informado por *corpus*. Temos ciência também de que nem todos se sentirão "maravilhados" com os resultados obtidos com a pesquisa em *corpus*, seja porque não "gostam de lidar com computador", porque "não têm tempo" para pesquisa, ou porque preferem "tudo pronto". Mas as duas oficinas-piloto confirmam o que McEnery (1998) afirmou há quase duas décadas: sem investir na capacitação docente para o uso de ferramentas da Linguística de Corpus, não adiantará nada publicar material baseado em *corpora*.

92

INTRODUCING BRAWE: A NEW RESOURCE FOR RESEARCHERS OF EAP

Larissa Goulart da Silva, Marine Laisa Matte e Simone Sarmento
Warwick University, Universidade Federal do Rio Grande do Sul (UFRGS)
l.goulart-da-silva@warwick.ac.uk, marine.laisa@gmail.com, simone.sarmento@ufrgs.br

The aim of this paper is to describe the compilation of a corpus of Brazilian academic English (BrAWE). This corpus allows other researchers to use it to investigate linguistic aspects related to English for Academic Purposes (EAP) produced by Brazilian students in comparison to other corpora *such as BAWE or MICUSP, for instance. The students represented in it are Brazilian students doing part of their undergraduate degree in British universities.*

95

INVENTIVIDADE LEXICAL NOS NOMES DE OPERAÇÕES POLICIAIS: UMA ANÁLISE POR MEIO DE UM *CORPUS* JORNALÍSTICO

Candice Guarato Santos
Universidade Federal de Uberlândia (UFU)
candiceguaratos@gmail.com

Este trabalho parte de duas questões: pode-se definir o estilo de nomear de uma instituição, no caso, da polícia? Há algum padrão nas denominações das operações? O

corpus de estudo foi compilado a partir de jornais, populares e de referência, de 11 cidades do estado de Minas Gerais. As teorias da Lexicologia, da Onomástica, da Estilística, dos Neologismos, e do Jornalismo, popular e de referência embasaram esta pesquisa. Os fundamentos da Linguística de Corpus (LC) colaboraram no desenvolvimento da metodologia. A pesquisa linguística, orientada pela LC, consiste na investigação de exemplos reais da língua retirados de textos autênticos que foram armazenados no meio eletrônico. Todo jornal, segundo Amaral (2011), é produzido em função de um determinado mercado, mesmo quando o lucro não seja pretendido. Há jornais que visam atingir as camadas sociais mais altas, enquanto há publicações que focam na audiência de menor poder aquisitivo. O primeiro tipo de jornal é chamado de jornal de referência, ou *quality paper* e o segundo tipo é denominado de jornal popular. Eles se diferem na forma de tratar o que é noticiado. O jornal popular noticia assuntos que fazem parte do cotidiano das classes mais baixas, além de apresentarem "preços baixos, planejamento gráfico atraente, linguagem acessível e anúncios de produtos e serviços voltados ao público de baixa renda" (Oliveira, 2009, p. 8). A credibilidade e o prestígio diante dos formadores de opinião são, de acordo com Amaral (2011), características imprescindíveis em um *quality paper*. No jornalismo de referência, o interesse pelo mundo público é primordial. Todos esses fatores são a causa dos *quality papers* respeitarem, segundo a jornalista, certos padrões éticos. A escolha por notícias tanto de jornais populares como de jornais de referência se justifica pela importância que cada um desses tipos de publicação tem na sociedade. Os jornais de referência possuem prestígio e servem de parâmetro para os formadores de opinião. Por outro lado, os jornais populares alcançam pessoas que até mesmo não tinham o hábito da leitura. Selecionamos os 11 jornais: Aqui, Correio de Uberlândia, Diário Popular, Estado de Minas, Folha de Contagem, Gazeta de Uberlândia, Jornal da Manhã, Jornal de Uberaba, O Tempo, Super Notícia e Tribuna de Minas. No total, foram compiladas 1.310 notícias sobre operações policiais. A pesquisa pelas notícias foi realizada por meio do Google, pois alguns *sites* dos jornais não possibilitavam buscas específicas. No espaço de busca do Google, inserimos as palavras operação e *site* que foram seguidas por dois pontos e pelo endereço eletrônico de cada jornal. As notícias foram arquivadas no formato txt e com a codificação unicode, como requerem as configurações do WordSmith Tools (Scott, 2012). As informações sobre a data de coleta, a fonte, o jornalista, autor da notícia, e a data de publicação foram inseridas entre chaves angulares. Esse é um recurso que impede o programa de processar certos dados. Cada arquivo foi nomeado de acordo com a data de publicação e armazenado em pastas homônimas aos seus jornais. Por meio da ferramenta Concord, buscamos pela palavra operação. A partir das linhas de concordância, identificamos os nomes das operações e acessamos o *source text* para identificarmos o sentido de cada operação. Em alguns casos, não havia a definição e nem o porquê da escolha do nome, pois não era a primeira vez que a ação policial foi noticiada. Em situações como essas, realizamos outra busca, no Google, por meio do comando "define: operação". Assim, a partir de outras notícias e foi possível identificar o sentido do nome da operação. Os nomes foram classificados de acordo com campos semânticos como, por exemplo, Datas/Feriados, Bens/dinheiro/patrimônio/propina, Doença/Hospital/Saúde, Ficção/Filme/Personagens, Grécia/Mitologia, Judiciário/Lei, Nome próprio de pessoa e Nome próprio de lugar. Por meio da análise, observamos que grande parte dos nomes das operações, presentes no *corpus*, são neologismos. O tipo mais comum é o neologismo semântico, seguido pelos empréstimos.

Entre os casos de neologismos semânticos, podemos destacar a Operação Genitora que prendeu um homem que chefiava um negócio do tráfico com seus irmãos. Segundo a notícia, o nome foi uma referência à mulher que deixou como herança para os três filhos o controle do tráfico de drogas das bocas de fumo. Os empréstimos mais comuns são oriundos do inglês (operações American Dream, Firewall, For All e Reset) seguido pelos provenientes do grego (operações Aletheia e Argos Panoptes). Identificamos denominações que são epônimos, ou seja, foram inspiradas em personagens de desenhos animados como a Operação Dick (nome inspirado no personagem Dick Vigarista), Operação Liga da Justiça e Operação Mandrake II. Filmes também inspiraram denominações como a Operação Krull e a Operação Triplo X. Os casos de nomes inspirados em topônimos são: Operação Carajás, uma alusão ao bairro de Contagem, Operação Faixa de Gaza, Operação Triângulo das Bermudas, Operação Samoa e Operação BH Segura. Acreditamos que os recursos linguísticos utilizados na criação dos nomes das operações, presentes no *corpus* de estudo, são empregados com a finalidade de causar algum efeito, como o de ressaltar a importância da ação policial, chamar a atenção por meio do humor ou por meio de nomes imponentes, uma vez que esses nomes circulam na imprensa.

97

A LINGUÍSTICA DE CORPUS EM INTERFACE COM ANÁLISE CONTRASTIVA E PRESSUPOSTOS DA TRADUÇÃO: UMA POSSIBILIDADE DE AUXÍLIO PARA PROMOÇÃO DE AUTONOMIA NA FORMAÇÃO DO PROFESSOR DE LÍNGUA INGLESA – ESTUDO DE CASO NO CAMPUS XIV DA UNEB

Lavine Cunha
Universidade Estadual da Bahia
lavinediilima@gmail.com

Neste ensaio trataremos de diferenças morfossintáticas entre Língua Portuguesa (LP) e Língua Inglesa (LI), especificamente *adjetival orders*. Através da abordagem da Análise Contrastiva (AC) (Fries e Lado, 1957), e pressupostos da tradução (Baker, 1992), em interface com a Linguística de Corpus mostraremos como o professor pode adquirir maior autonomia durante aquisição de LI quanto ao comportamento dos adjetivos.

110

ELABORAÇÃO DE UM PROTÓTIPO DE GLOSSÁRIO BILÍNGUE (PORTUGUÊS-INGLÊS) DE TREINAMENTO DE FORÇA: SUBSÍDIOS PARA O TRADUTOR

Márcia dos Santos Dornelles
Universidade Federal do Rio Grande do Sul (UFRGS)
marcia@esef.ufrgs.br

O terminógrafo, ao elaborar um produto terminográfico bilíngue baseado em *corpus* para tradutores, deve preocupar-se não só em repertoriar, nas duas línguas, os termos

próprios de uma (sub)área do conhecimento, mas também em apresentá-los inseridos em suas combinatórias típicas, ou seja, associados aos elementos que a eles se combinam em nível sintagmático, de forma recorrente nos textos daquela especialidade. Isso porque o tradutor precisa produzir um texto de chegada adequado ao padrão de linguagem em foco, de forma a espelhar o *modus dicendi* daquele campo. Assim, seu texto soará natural à comunidade de leitores, evitando-se ruídos na comunicação. Para tanto, assim como um biólogo precisa explorar o meio em que vive seu espécime de estudo para entender o comportamento deste, também o terminógrafo precisa conhecer o *habitat* dos termos: o texto especializado. Nesse sentido, o conhecimento das propriedades do gênero textual em estudo qualifica um produto terminográfico, considerando que os termos e demais elementos a ele incorporados, como as fraseologias especializadas, os contextos definitórios e os exemplos de uso, extraídos de seu âmbito natural de emprego, ajudam a compor os modos de dizer desse gênero. Somente dessa forma um produto terminográfico tem chances de ser aceito pela comunidade da área. Em síntese, o gênero textual é um elemento condicionante do perfil das terminologias. Com esses pressupostos e diante da falta de produtos terminográficos bilíngues no âmbito do Treinamento de Força, especialmente dirigido a tradutores brasileiros, esta pesquisa de mestrado (Dornelles, 2015b) teve como objetivo central apresentar bases teórico-metodológicas consistentes para a elaboração de um glossário de Treinamento de Força na direção português→inglês, destinado especialmente a tradutores, mas útil também para pesquisadores e estudantes dessa temática que precisem produzir artigos científicos em inglês. Os objetivos secundários, todos alcançados, foram (a) oferecer um protótipo do glossário, composto de guia do usuário, uma árvore de domínio em português do Treinamento de Força, lista de termos em português e 30 exemplares de fichas terminológicas em formato estendido; e (b) oferecer uma descrição do comportamento das unidades terminológicas em português e inglês, e das unidades fraseológicas especializadas (UFE) eventivas (Bevilacqua, 2003; 2004) em português nos artigos sobre Treinamento de Força. Como referencial teórico, valemo-nos dos princípios da Teoria Comunicativa da Terminologia (TCT) e dos fundamentos e diretrizes da Linguística de Corpus (LC). Seguir a TCT (Cabré, 1999a; 1999b; 2001a; 2001b; 2003; 2009) implica adotar o termo como objeto central de estudo e concebê-lo, antes de tudo, como uma unidade lexical da língua natural que adquire valor especializado dentro de um contexto especializado, segundo critérios semânticos, discursivos e pragmáticos. Seguir a LC (Berber Sardinha, 2004; Biber, 2012) implica uma visão probabilística da língua, pressupondo que, embora muitos traços linguísticos sejam possíveis teoricamente, não ocorrem com a mesma frequência. Ganham, ainda, realce no estudo os temas do artigo científico como gênero textual especializado; da variação terminológica (Freixa, 2002; 2014); da tradução como processo (Hurtado Albir, 2008); da competência tradutória (Pacte, 2011); e da equivalência funcional (Gémar, 1998; Nord; 1998). Nosso *corpus* de estudo é constituído de 70 artigos – originais e de revisão – de periódicos científicos de destaque no âmbito do Treinamento de Força, escritos originalmente em português e inglês. São, portanto, dois *subcorpora*, um em cada língua, que são comparáveis. Os periódicos, todos online, foram recomendados por um consultor especialista em Treinamento de Força. A busca inicial dos artigos foi feita nas páginas eletrônicas das revistas na base de dados SciELO (artigos em português) e no Portal de Periódicos da Capes (revistas em inglês), pelos assuntos "treinamento de força" e "*strength training*", respectivamente. Nas listas de arti-

gos resultantes por revista, selecionamos aqueles publicados a partir de 2002. Após, com o auxílio do consultor, excluímos os textos que pareciam descolar o tema do Treinamento de Força do âmbito da Educação Física, ou seja, não focavam o treinamento em si. Eliminamos, assim, artigos de pesquisas com animais; artigos com enfoque clínico; e aqueles mais inseridos no campo da Bioquímica. Para exploração e análise do *corpus*, utilizamos o software AntConc (Anthony, 2011), especialmente as funcionalidades *keyword list*, *n-grams* e *concordance*. Como material de apoio, utilizamos livros-texto e artigos científicos de referência sobre Treinamento de Força, um glossário particular preexistente de Educação Física, a Terminologia Anatômica Internacional (SBA, 2001), o Google Acadêmico, o Wikipédia, entre outros. Os critérios para o fichamento das unidades terminológicas foram a distribuição e a frequência no *subcorpus* em português; a pertinência temática e a pertinência pragmática (Maciel, 2001); o encaixe na árvore de domínio; e a existência de pelo menos um equivalente em inglês, preferencialmente no *subcorpus* em inglês. Como resultados da investigação, oferecemos um protótipo de glossário composto de árvore de domínio em português, com uma população de 71 unidades terminológicas (UT) e uma amostra fichada de 30 UT (42,25%); Guia do usuário; uma Lista de termos em português, com 30 UT-lema e 89 UT variantes; 30 fichas terminológicas; e 78 termos em inglês, sendo 30 equivalentes preferenciais e 48 variantes. A microestrutura da ficha terminológica, baseada nas propostas de Fromm (2007) e Teixeira (2008), contempla os seguintes itens: unidade terminológica (UT) em português; sigla/abreviatura/acrônimo/fórmula/símbolo, conforme o caso; informação gramatical; frequência da UT no *corpus*; área e (sub)domínio; número de posição na árvore de domínio (com *hiperlink* para a árvore); figura; *link* para vídeo (p. ex., para demonstrar um exercício); definição simplificada em português; outras definições, com as fontes; variante(s) em português, com remissiva (*hiperlink*), frequência no *corpus* e observações sobre uso; equivalente(s) em inglês com a frequência e observações sobre uso; fraseologia(s) em português; equivalente(s) em inglês das fraseologias; exemplo(s) de ocorrências em português e inglês; UT relacionadas, com remissivas (*hiperlink*) para as respectivas fichas; notas de uso; notas de tradução; e dados de revisão da ficha. Como características da terminologia do Treinamento de Força em artigos científicos, quanto à sua morfossintaxe, verificamos um número bastante mais elevado de UT polilexicais (85% em português e 86% em inglês) em relação às monolexicais (9 e 8%) e às siglas/abreviaturas (6% nas duas línguas). As estruturas mais recorrentes nos termos em português foram quatro: N + prep (+ art) + N (34%), p. ex., intervalo de recuperação; N + ADJ (30%): pesos livres; N + N (6%): rosca bíceps; N + ADJ + ADJ (6%): força máxima isométrica. Em inglês, as estruturas mais recorrentes foram cinco: N + N (35%): *strength training*; ADJ + N (18%): *free weights*; ADJ + N + N (8%): *single-joint exercises*; N + V (6%): *biceps curl*; ADJ + ADJ + N (5%): *maximum isometric force*. As UFE eventivas em português não foram numerosas como esperávamos, mesmo adotando uma frequência/distribuição não muito alta, como é a 2/2. Por esse critério, encontramos 33 unidades no *corpus* de estudo. As ferramentas do AntConc mais utilizadas para o seu reconhecimento foram as *concordances*, seguidas dos *clusters*. Os núcleos eventivos com nominalizações foram três vezes mais frequentes que com verbos e quatro vezes mais frequentes que com particípios. Veja-se: nominalizações: 21 UFE (64%), como em prática de treinamento de força, combinação {de/das} variáveis do treinamento, aumento {da/na} intensidade do treinamento, execução {de/dos} exercícios de força; verbos: 7 UFE (21%), como realizar

(um) treinamento de força, determinar a intensidade do treinamento, executar [NUM] séries, etc.; e particípio: 5 UFE (15%), como treinamento de força realizado, número de repetições completadas, unidades motoras recrutadas. Há certo grau de variação (comutação) também nos núcleos eventivos das UFE, o que não afeta sua estabilidade semântica nem as descaracteriza como unidades semifixas. Certos núcleos eventivos das UFE carregam um valor terminológico que se equipara ao de seus núcleos: p. ex., recrutamento/sincronização de unidades motoras. Destaca-se, ainda, nos resultados desta pesquisa e de outros três estudos exploratórios anteriores (Dornelles, 2014a; 2014b; 2015a) uma expressiva variação terminológica intra e intertextual, nas duas línguas. Algumas UT apresentam, ao mesmo tempo, variação denominativa e uma sutil variação conceitual. Em português, cada UT-lema apresentou de 0 a 7 variantes. Quanto aos tipos de variação (cf. Freixa, 2002), encontramos a lexical: 31 variantes (53%), como nos pares treinamento de força/treino de força; exercícios de força/exercícios resistidos; extensão de joelho/extensão de perna(s); por redução: 12 (20%), como em rosca scott/rosca bíceps scott; barra livre/barra; hipertrofia muscular/hipertrofia; morfossintática: 7 (12%); gráfica: 7 (12%), como em volume do treinamento/volume de treinamento; extensão de joelho/extensão de joelhos; força máxima isométrica/força isométrica máxima; e complexa (lexical e redução concomitantemente): 2 (3%), em treinamento de força/musculação e treinamento de força/treino resistido. Em inglês, cada equivalente preferencial apresentou de 0 a 5 variantes. Os tipos foram lexical: 21 variantes (43%), como nos pares *strength training*/*resistance training*; *rest period*(*s*)/*rest interval*(*s*); gráfica: 11 (23%), como em *repetition*(*s*)/*rep*(*s*); *repetition maximum*/RM; *multi-joint exercise*(*s*)/*multijoint exercise*(*s*); por redução: 8 (17%), como em *training variables*/*training program variables*; *concentration curl*/*biceps concentration curl*; e morfossintática: 8 (17%), como em *rest period*(*s*)/*resting period*(*s*); *muscle power*/*muscular power*. A pesquisa contempla, então, uma parte teórica e uma parte aplicada que se inter-relacionam e se inserem na dupla face da Terminologia, visto que há uma descrição de uma linguagem especializada a partir de um dado ponto de vista teórico e o desenho de um produto concreto.

Resumos de trabalhos em andamento “Minuto de Loucura”

AELIUS FALADO 2.0: PRIMEIROS PASSOS

Mônica Rigo Ayres e Gabriel de Ávila Othero
Instituto de Letras da Universidade Federal do Rio Grande do Sul (UFRGS)
monicarigoayres@hotmail.com

Este trabalho insere-se nas áreas de Linguística Computacional, Linguística de Corpus e análise morfossintática do português brasileiro, mais especificamente sobre anotação automática morfológica de *corpora* de língua falada. Pretendemos propor uma nova versão do etiquetador automático morfossintático Aelius, desenvolvido originalmente pelo prof. dr. Leonel Alencar, da Universidade Federal do Ceará, coordenador do projeto Computação e Linguagem Natural (CompLin). Esse etiquetador foi planejado e desenvolvido para etiquetar textos de língua escrita, e cumpre seu papel satisfatoriamente. A nova versão que estamos desenvolvendo será desenhada para etiquetar textos de língua falada. As etiquetas utilizadas no Aelius foram propostas pelo desenvolvedor do programa, e são as mesmas do *corpus* de português histórico Tycho-Brahe (*corpus* eletrônico já anotado, composto de textos em português). Os textos que utilizamos para a primeira fase do processo de anotação de língua falada são do banco do projeto Varsul, que estuda a variação linguística na região sul do Brasil.

Em trabalhos anteriores (Ayres, 2014; Othero e Ayres, 2014), verificamos o desempenho do Aelius na etiquetagem morfossintática automática de um trecho do *corpus* do Varsul (constituído por 20 amostras de textos, num total de 154.530 palavras), um desafio inédito para a ferramenta até então. A partir dos erros do etiquetador, buscamos depreender certos padrões de anotação para superar limitações apresentadas pelo programa, propondo algumas sugestões de implementações para que o Aelius etiquetasse de maneira mais satisfatória *corpora* de língua falada. Tratamos especialmente dos casos de interjeições, aféreses, onomatopeias e marcadores discursivos/conversacionais, que são marcas específicas da oralidade.

Ao longo dos últimos anos, a Linguística de Corpus no Brasil aumentou seu escopo de pesquisa, abrangendo a língua falada além da língua escrita (Berber Sardinha, 2000), e as transcrições de *corpora* falados possuem características específicas da fala, tais como repetições, contrações, aféreses, marcadores conversacionais, fragmentos de sentenças, interjeições e truncamentos. Por isso, é de extrema importância que o anotador dê conta de etiquetar corretamente todas essas informações que fazem parte da fala (como no trecho seguinte).

Não/NEG sei/VB-P o/D quê/WPRO ,/, aí/ADV eu/PRO disse/VB-D :/. “/QT Ah/INTJ ,/, que/C delícia/N isso/DEM !/. Que/C bom/ADJ é/SR-P gente/N da/P+D-F minha/PRO$-F terra/N ,/, né/NPR ?/. Que/C está/ET-P indo/VB-G aqui/ADV “/QT ./. Muito/Q gaúcho/N ,/, claro/ADJ ,/, né/NPR ?/.

A palavra “claro”, nesse contexto, não é um adjetivo, como foi analisada pelo etiquetador, mas um marcador conversacional. Como essa é uma palavra lexical (diferentemente de outros marcadores conversacionais, como “hum”, e “tarará”, por exemplo – que também apareceram no *corpus* que analisamos), ela tem diferentes possibilidades de significado, i.e., dependendo do caso deve receber uma etiqueta diferente; por isso, o algoritmo precisa ser bem treinado. Sendo assim, deve-se levar em conta

a posição da palavra na frase, por exemplo: se a palavra "claro" aparece imediatamente após um nome, a chance de ser um adjetivo é muito grande, mas se aparece entre vírgulas, a probabilidade é de que se trata de um marcador conversacional. Pretendemos buscar soluções para que uma nova versão do etiquetador consiga lidar satisfatoriamente com ocorrências como essas, específicas da oralidade.

15

ALIMENTOS REGIONAIS BRASILEIROS: ESTATÍSTICA LEXICAL, GÊNERO TEXTUAL E ACESSIBILIDADE TEXTUAL

Laura Dupke de Almeida
Universidade Federal do Rio Grande do Sul (UFRGS)
lauradupke@gmail.com

Com apoio de recursos da Linguística de Corpus (Berber Sardinha, 2004), explora-se aqui, em um estudo inicial, o perfil do vocabulário da segunda edição do livro *Alimentos regionais brasileiros*, um guia de 486 páginas do Ministério da Saúde do Brasil, lançado em 2015, que visou divulgar a alimentação saudável para qualquer cidadão interessado no tema. Tomado esse livro como um *corpus* de estudo, busca-se reconhecer o gênero textual nele envolvido conforme Marcuschi (2002) e identificar o perfil de vocabulário partindo-se dos trabalhos de Evers (2013) e de Finatto (2014). O objetivo do trabalho é realizar a análise dessa obra em particular para, futuramente, somá-la a um acervo de produções escritas, de diferentes gêneros, destinadas a leitores de escolaridade limitada e pouco hábito de leitura. Com os programas AntConc, Flesch Calculator e TermoStat Web, observam-se, em uma amostra de 40 páginas do mesmo, a lista e a frequência de palavras, Índice Flesch (IF), *types*/*tokens ratio* (TTR), trigramas mais frequentes e lista de *keywords*, características ponderadas com a função da obra e seu público-alvo. A questão da acessibilidade textual, considerando elementos como escolha de palavras e uso de vocabulário específico também é ponderada a partir dessa descrição inicial da obra. Embora em fase inicial, preliminarmente, a partir da análise da amostra, com 5.749 *tokens* e 1.306 *types*, já se pode perceber: a) vocabulário relativamente repetitivo, com TTR de 20,04% – pois, em contraste, um artigo científico de Pediatria tende a exibir 24% de TTR (cf. Finatto, 2011); b) texto de leitura muito fácil, com verificação do IF em 93,83, enquanto um artigo de Pediatria mostra cerca de 27 (Finatto, 2011) sendo bastante difícil; c) "de" e "e" como palavras gramaticais mais frequentes – seguindo o mesmo padrão de diferentes textos escritos em português (Zilio, 2010; Finatto, 2011 e 2014); quanto às palavras lexicais, as mais usadas são abreviaturas, tais como "g" (grama), "mg" (miligrama), entre outros. Por tratar sobre alimentos regionais do Brasil, há vários nomes de ingredientes que poderiam causar estranhamento para leitores de determinadas localidades, como "abiu", "camu-camu", "abricó", etc. Com essa amostra do livro, pressupõe-se que o texto pretenda integrar público leigo e especialistas, pois traz diferentes tipos de informações técnicas sobre alimentos regionais, assim como receitas e possíveis preparações culinárias com os alimentos apresentados. Assim, ainda que se trate aqui de explorar uma única obra, poder-se-ia identificar, a princípio, dois gêneros textuais envolvidos: 1) enciclopédia, pois apresenta uma série de alimentos separados pela região proveniente,

e um conhecimento aprofundado, técnico-científico, sobre cada um; 2) receita culinária, com ingredientes e modo de fazer. Esses dois gêneros aparecem intercalados, sendo que o primeiro, em tese, dada a presença de terminologias sem a devida explicação, poderá trazer problemas de acessibilidade para leitores de escolaridade limitada. Outro aspecto a salientar é o uso vocabulário regional. Este estudo deve prosseguir com foco específico na parte enciclopédica, objetivando descrever possíveis pontos de complexidade do vocabulário técnico-científico – especialmente da terminologia – para leitores leigos. Feito isso, também serão tratados alguns elementos peculiares dos vocabulários regionais na parte das receitas, que, teoricamente, deveriam ser compreensíveis para qualquer leitor do Brasil. Para o estudo das receitas e sua complexidade, pretende-se recorrer ao trabalho de Teixeira (2008). Apoios: PROPESQ-UFRGS e CNPq.

16

A ELABORAÇÃO DE UM DICIONÁRIO DE ELEMENTOS COESIVOS COM BASE EM *CORPUS*

Daniela Faria Grama e Guilherme Fromm
Universidade Federal de Uberlândia (UFU)
daniela_grama@hotmail.com

Este resumo refere-se a uma pesquisa de Doutorado que está em fase inicial e que visa à elaboração de um dicionário de elementos de coesão do português, com base em *corpus*, para alunos do Ensino Médio. Para realização da pesquisa, prevemos um trabalho metodológico minucioso com o *corpus* de redações que constituímos no Mestrado, conforme Grama (2016). Inicialmente, aumentaremos tal *corpus*, que atualmente possui 1.399 textos do tipo dissertativo-argumentativo, com as redações de 2015 e 2016 disponibilizadas no site UOL Educação. Depois, revisaremos o *corpus* de redações em sua totalidade para verificar se restaram palavras unidas e colchetes com sugestões de correção. Para isso, utilizaremos as ferramentas WordList e Concord do WordSmith Tools (Scott, 2012). Outra etapa importante compreende todo o processo de descrição e análise do *corpus* de redações que será feito com o intuito de obtermos as palavras que farão parte da macroestrutura do dicionário e as informações que constarão na estrutura da ficha lexicográfica, produzida no mestrado, de cada palavra-entrada do dicionário de elementos coesivos. Utilizaremos o *corpus* para verificar quais elementos coesivos (unipalavras e multipalavras) existem e devem fazer parte da macroestrutura do dicionário. Para isso, faremos uma *wordList* do *corpus* e anotaremos as palavras candidatas a elementos coesivos. Verificaremos o contexto linguístico de cada uma, por meio da ferramenta Concord, a fim de confirmarmos se são elementos de coesão. Vale ressaltar que esse procedimento será realizado apenas com aquelas palavras que não fazem parte da lista de 255 elementos de coesão que fizemos no mestrado. Também usaremos o *corpus* para identificar a frequência das palavras na função de elementos coesivos. Devido à potencialidade que as palavras da língua portuguesa possuem, é preciso verificar os seus contextos linguísticos, por meio da ferramenta Concord, pois, como vimos em nossa dissertação, nem sempre "porém", por exemplo, será usado como conjunção, ele pode ser usado como substantivo. Portanto, esse procedimento é fundamental para

obtermos a frequência exata de uso de cada elemento coesivo. Além disso, usaremos o *corpus* para contabilizar a posição textual que cada elemento coesivo é utilizado e para analisar o comportamento/uso dos elementos coesivos, o que nos auxiliará a elaborar a definição e as notas (contaremos também com as contribuições da Linguística Textual e da Semântica Argumentativa). O *corpus* de redações será útil para extrairmos exemplos de uso. Para isso, utilizaremos os mesmos critérios estipulados em nossa dissertação de mestrado: exemplos que estejam em consonância com a definição elaborada, que levem em consideração não apenas a dimensão frasal, mas também a textual, que tenham coerência e estrutura sintática majoritariamente bem formada e que respeitem os direitos humanos. Temos o intuito de acrescentar no Back Matter do dicionário, que será uma ferramenta de consulta online, exercícios que possam auxiliar os consulentes a compreenderem a importância do uso dos elementos de coesão em uma produção textual. Provavelmente, o *corpus* será útil nessa tarefa também. Assim, com o auxílio do *corpus* de redações, esperamos obter resultados satisfatórios que contribuirão para a elaboração do nosso dicionário de elementos coesivos.

19

ORGANIZAÇÃO TEMÁTICA DE FOLHETOS INFORMATIVOS PARA AGRICULTORES FAMILIARES DO BRASIL: UM ESTUDO BASEADO EM *CORPUS*

Giselle Fetter
Universidade Federal do Rio Grande do Sul (UFRGS)
gisellerevisora@gmail.com

A partir de estudos sobre Acessibilidade Textual (Bailin e Grafstein, 2016; Myers, 2002, 2004; Coulthard, 2004; Leffa, 1996), com ênfase no gênero divulgação científica (Finatto et al., 2016; Ciapuscio, 2003) e da Linguística Sistêmico-Funcional (Halliday e Matthiessen, 2014; Fuzer e Cabral, 2014), este trabalho, com apoio de recursos da Linguística de Corpus, tem como foco a terminologia presente nas estruturas temáticas dos textos de folhetos informativos institucionais direcionados aos agricultores familiares. O objetivo da descrição é subsidiar uma avaliação terminológica com vistas à acessibilidade textual desses materiais para leitores de escolaridade limitada. Conforme a perspectiva sistêmico-funcional, o texto é determinado pelo contexto, em outras palavras, a gramática de uma língua está em conexão com o seu entorno e com os processos sociais em que os indivíduos estão envolvidos (Halliday e Hasan, 1989). Assim, as escolhas lexicogramaticais do indivíduo se manifestam conforme seu propósito e podem ser analisadas a partir de três metafunções: ideacional, representação de mundo interno ou externo; interpessoal, expressão da relação entre as pessoas; e textual, a oração como mensagem. Esta pesquisa, ao descrever o texto escrito desses folhetos, destacará a metafunção textual, em que a oração é observada por suas duas partes constitutivas denominadas Tema e Rema. O Tema determina o contexto da oração e é escolhido pelo emissor para orientar o receptor. Já o Rema é responsável pelo desenvolvimento do Tema. O *corpus* de estudo é composto por 30 folhetos produzidos pela Associação Riograndense de Empreendimentos de Assistência Técnica e Extensão Rural (Emater-

-RS) e, como *corpus* de contraste, tem-se 30 folhetos produzidos pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa). Primeiramente, realizou-se uma anotação manual dos tipos de Temas das frases do *corpus* e seu levantamento lexical-quantitativo com o recurso Concord da ferramenta WordSmith Tools. Com o auxílio dessa ferramenta, foram analisadas 4.867 orações de folhetos da Emater/RS e da Embrapa, excetuando-se as frases cujos Temas não puderam ser definidos. Os resultados iniciais apontam que folhetos da Emater/RS apresentam: a) maior ocorrência de Temas Simples; b) frases que não retomam o Tema da frase anterior, o que pode prejudicar a continuidade da leitura; c) alta incidência de termos científicos; d) termos presentes nos Temas sem definição; e) pouco uso de elementos textuais como conjunções e adjuntos conjuntivos. Os folhetos da Embrapa apresentam: a) alta incidência de elementos textuais; b) alta incidência de elementos circunstanciais, que podem colaborar para o desenvolvimento da mensagem; c) uso de elementos interpessoais que expressam opinião. Na continuidade desta pesquisa, pretende-se apontar a total representatividade dos tipos de Temas nos folhetos de ambas instituições, o total de termos presentes nas estruturas temáticas e, por fim, verificar qual a estratégia, no caso de sua ocorrência, utilizada nos folhetos para a definição desses termos. Também se pretende, após a obtenção dos resultados finais, fornecer subsídios para a escrita de materiais instrucionais mais acessíveis para leitores com perfil compatível de nossos agricultores familiares.

21

SOFTWARE PARA ALFABETIZAÇÃO DE AUTISTAS NO MUNDO DA ASTRONOMIA

Guilherme Queiroz Vasconcelos, Cinthyan Renata Sachs C. de Barbosa, João Coelho Neto e Fernanda Mafort
Universidade Estadual de Londrina; Universidade Estadual do Norte do Paraná – Cornélio Procópio; Pontifícia Universidade Católica do Paraná
gui.queirozv@gmail.com; cinthyan@uel.br; joaocoelho@uenp.edu.br; fernanda-mafort@hotmail.com

O presente trabalho apresenta os passos iniciais do projeto de um software para alfabetização de crianças autistas, traçando um aporte teórico e os passos metodológicos utilizados para construção do software, apresentado o conteúdo de alfabetização desenvolvido, bem como as propostas futuras.

METÁFORAS SOBRE REFUGIADOS NO JORNAL *FOLHA DE S. PAULO* EM 1994 E 1995

Luciane Corrêa Ferreira e Catarina Valle e Flister
Universidade Federal de Minas Gerais (UFMG)
lucianeufmg@gmail.com, catarinaflister@gmail.com

Essa pesquisa busca descobrir quais são as metáforas relacionadas aos itens lexicais *refugiado*, *refugiados*, *refugiada*, *refugiadas* no jornal diário Folha de S. Paulo nos anos de 1994 e 1995. Trabalhamos com 730 edições do jornal referido, disponibilizados através do site da Linguateca, na coleção CHAVE, mais especificamente na CHAVEFolha. Analisamos as notícias a partir do quadro teórico da Linguística Cognitiva. Dessa forma, trabalhamos com a Teoria da Metáfora Conceptual de Lakoff e Johnson (1980), que postula que as metáforas, embora expressadas pela linguagem, não são figuras de linguagem, mas sim figuras de pensamento. A partir da Teoria da Metáfora Conceptual, buscamos descobrir a ideologia subjacente acerca dos temas dos refugiados no jornal pesquisado. A ideia de que podemos detectar ideologias através da identificação da MC no discurso está em Charteris-Black (2004). Esse autor afirma que a metáfora, além de linguística e cognitiva, é pragmática. Assim, ela deve ser investigada considerando o contexto de produção e as intenções do falante – nesse caso, as intenções do meio de comunicação. O tema dos refugiados foi selecionado para nossa pesquisa por se tratar de um tema atual na agenda midiática mundial. Escolhemos a década de 1990 porque essa é uma pesquisa inicial, a partir da qual pretendemos conduzir outras. A década de 1990 marcou o início de uma política mais favorável aos refugiados no Brasil. Assim, achamos importante mapear a ideologia vigente a partir desse momento. Além disso, a questão dos refugiados é socialmente relevante e presente, o que faz com que a linguagem relacionada ao tema esteja em constante evolução, visto que novos significados estão sempre sendo construídos à medida em que novos pensamentos são comunicados. Para conduzir a análise dos dados, usamos a metodologia da Linguística de Corpus. Especificamente, usamos Stefanowitsch (2006), que pesquisa a metáfora no *corpus*. Para explicar tal metodologia, é necessária uma breve explicação acerca das nomenclaturas usadas na Teoria da Metáfora Conceptual. A metáfora consiste em "tomar uma coisa em termos de outra" (Lakoff e Johnson, 1980, p. XX), ou seja, é uma relação entre dois conceitos. Na teoria de Lakoff e Johnson, cada conceito ocupa o lugar de um domínio. O conceito usado enquanto metáfora é o domínio alvo e o conceito usado para entender o outro domínio conceptual é o domínio fonte. O método de Stefanowitsch (2006) consiste em selecionar um item lexical do domínio alvo, buscar as ocorrências desse item, identificar as expressões metafóricas que aparecem nesse resultado – a partir da definição de Lakoff e Johnson (1980) – e, finalmente, em agrupar mapeamentos semelhantes. Esse método nos permite visualizar os *frames* discursivos metafóricos que subjazem os discursos jornalísticos. A análise total das notícias ainda não foi completada. No entanto, podemos indicar como resultado parcial a presença da metáfora REFUGIADO É CRIMINOSO que aponta para a ideologia dominante REFUGIADO É RUIM. Tal resultado pode ser uma evidência forte de que os refugiados não são bem-vindos no Brasil.

PLANEJAMENTO, COMPILAÇÃO E ALINHAMENTO DE UM *CORPUS* PARALELO DE LITERATURA CANADENSE: TRADUÇÃO E RETRADUÇÃO EM PORTUGUÊS BRASILEIRO DE *LIFE OF PI*

Raphael Marco Oliveira Carneiro
Universidade Federal de Uberlândia (UFU)
raphael.olic@gmail.com

Dado o crescente número de pesquisas nos Estudos da Tradução que tem tratado de retraduções, este estudo busca contribuir para as discussões dentro desse tema. Tomando a primeira década do século XXI, vários estudos confirmaram ou refutaram a Hipótese da Retradução como proposta por Antoine Berman (Koskinen e Paloposki, 2010). Esses estudos também reconheceram a necessidade de mais pesquisas para ampliar o entendimento desse fenômeno complexo. Assim, esta investigação é uma tentativa inicial de testar essa hipótese em um *corpus* paralelo bilíngue constituído pelo romance *Life of Pi*, do autor canadense Yann Martel, pela tradução *A Vida de Pi* da tradutora brasileira Alda Porto e pela retradução *A Vida de Pi* da tradutora brasileira Maria Helena Rouanet.

É preciso definir que as traduções feitas após uma primeira tradução são chamadas de retraduções (Berman, 1990). Considera-se que uma retradução é uma segunda tradução na mesma língua da primeira tradução, e assim sucessivamente, haja vista quantas retraduções forem feitas de um mesmo texto-fonte.

Para explicar a necessidade de se retraduzir, Berman (1990) entende que as retraduções são produzidas porque as primeiras traduções envelhecem e são percebidas como defeituosas e, por isso, carentes de aprimoramentos. Os melhoramentos seriam, então, responsabilidade das retraduções, que se tornam cada vez mais semelhantes ao texto original, à medida que novas retraduções são produzidas. Essas considerações vieram a ser compreendidas como uma hipótese, a Hipótese da Retradução (HR), que pode ser assim formulada: a retradução de um texto X tende a se assemelhar ao texto X e a se afastar da tradução desse mesmo texto X.

O texto-fonte em análise nesta pesquisa foi primeiramente publicado no Canadá em 2001 e ambas as traduções foram publicadas no Brasil por editoras diferentes em diferentes anos, a primeira em 2004 e a última em 2010. No evento da adaptação fílmica, uma nova edição foi lançada em 2012 com o título *As Aventuras de Pi* com a mesma tradução da edição de 2010. *Life of Pi* é possivelmente um dos romances canadenses mais populares já traduzidos no Brasil. Uma consulta rápida a sítios de compras na *web* revela que é a obra mais comentada pelos leitores brasileiros em comparação a outros autores canadenses contemporâneos internacionalmente reconhecidos, como Alice Munro, Carol Shields, Margaret Atwood, Michael Ondaatje e Rohinton Mistry. É tanto por essa popularidade quanto pela publicação de uma nova tradução que nos interessou desenvolver esta pesquisa. O *corpus* foi inicialmente planejado como uma amostra de textos literários não anotada sendo um texto-fonte em inglês canadense e duas traduções em português brasileiro. Até onde pudemos atestar, esse *corpus* ainda não foi investigado no âmbito dos Estudos da Tradução.

Esta pesquisa parte dos conceitos e métodos dos Estudos Descritivos da Tradução

(EDT) e dos Estudos da Tradução Baseados em Corpus (ETBC), de modo que consideramos a descrição funcional de evidências empíricas compiladas como um *corpus* eletrônico. Destacamos o uso a ser feito dos conceitos de adequação e aceitabilidade provenientes dos EDT (Toury, 2012) ao tratar das relações entre o texto-fonte e as duas traduções. Depois de compilados, os dados do *corpus*, tanto de natureza estatística quanto de natureza linguística, foram obtidos por meio das ferramentas Concord e WordList do programa WordSmith Tools 6.0. Para o alinhamento do texto-fonte com os dois textos-alvo utilizamos o programa ParaConc.

Antes de partir para a análise do *corpus* como um todo, realizamos um estudo preliminar apenas com uma porção dos textos, ou seja, ***Author's Note/Nota do autor***, que se configura como um tipo de prefácio, antecedendo o início do romance. Características estilísticas das duas traduções foram tanto quantitativa quanto qualitativamente observadas em relação à razão forma/item, à média do tamanho das frases, às escolhas lexicais e à apresentação do discurso, em comparação ao texto-fonte. Essas características foram usadas como indicadores preliminares para verificar se a retradução é estilisticamente mais semelhante ao texto-fonte do que a tradução. Neste estágio inicial não podemos asseverar se a hipótese anteriormente mencionada será confirmada ou não. Porém, essa análise inicial indicou como as tradutoras lidaram com diferentes aspectos do texto, como na tradução de unidades lexicais estrangeiras. Também revelou que a nota do autor é um elemento paratextual importante para a interpretação e recepção do romance, assim como para a construção do mundo ficcional engendrado pelo texto.

32

TERMINOLOGIA EMPRESARIAL: O RELATÓRIO DE SUSTENTABILIDADE E AS LINGUAGENS ESPECIALIZADAS

Carolina R. Ourique
Universidade Federal do Rio Grande do Sul (UFRGS)
carolinacamargo79@gmail.com

As empresas dos mais diversos segmentos de produção que utilizam recursos naturais publicam relatórios de sustentabilidade com o propósito de demonstrar aos públicos de interesse o impacto de suas atividades no meio ambiente. O objetivo deste trabalho é caracterizar esse relatório como um texto de linguagem especializada no qual o discurso se organiza em torno da terminologia prototípica de uma área de conhecimento e atividade. Nesse direcionamento, busca-se apoio nos referenciais teóricos da Terminologia de Perspectiva Comunicativa (Cabré, 1999), para a qual o texto especializado é considerado uma realização da língua natural com características próprias de uma área temática, do objetivo e do modo de expressão dos seus especialistas. Além disso, recorre-se aos princípios teórico-metodológicos de perspectiva textual (Ciapuscio, 1993) que consideram os textos como objetos complexos que incluem diferentes níveis de análise: funcional, situacional, temático e linguístico. Adota-se também a abordagem da Linguística de Corpus (Berber Sardinha, 2004) que preconiza a busca do significado no contexto real da comunicação. Nesse quadro referencial, são analisadas amostras de diferentes relatórios em português e inglês, tendo como base os conceitos de adequação da perspectiva funcio-

nalista dos atuais Estudos de Tradução (Nord, 2009), segundo os quais não se traduzem elementos estruturais isolados de um texto, mas funções comunicativas dependentes do sistema linguístico, das normas e convenções específicas de cada cultura. Foi construído um *corpus* paralelo composto por três textos de empresas brasileiras redigidos e publicados originalmente em português, alinhados paralelamente com suas respectivas traduções para o inglês. Os textos foram selecionados pelos seguintes critérios: constar no Guia de Sustentabilidade (Revista Exame, 2015) como empresas mais sustentáveis em seus respectivos setores, estar em português e estar disponibilizados online ao público em geral. Para iniciar, foram destacados e examinados manualmente os elementos constitutivos da macroestrutura de cada documento. Em seguida, com assistência das ferramentas informatizadas do software de análise textual AntConc (Anthony, 2014), foram analisados os padrões léxicos, morfossintáticos e pragmáticos através das evidências fornecidas pelas listas de concordâncias das palavras mais frequentes no *corpus*, pelo levantamento dos agrupamentos lexicais mais significativos e pelo exame dos dados no contexto. Nesta fase inicial do projeto, foi possível identificar alguns dos elementos que configuram o caráter especializado dessas publicações. Com os resultados desse estudo exploratório, espera-se aprofundar a identificação dos elementos que configuram a especialização desse tipo textual e assim contribuir ao melhor desempenho de redatores e também de tradutores empresariais.

33

VOCABULÁRIO BILÍNGUE PORTUGUÊS/INGLÊS DE LINGUÍSTICA GERAL: PADRONIZAÇÃO DOS *CORPORA*

Márcio Issamu Yamamoto
Universidade Federal de Goiás (UFG)
issamu2009@gmail.com

A Linguística passa a ser disciplina ministrada nos cursos de Letras das universidades brasileiras (Bagno e Rangel, 2005; Castilho, 2007) a partir dos anos 1960. E, atualmente, há grandes centros que se dedicam ao estudo desta área no Brasil, com produções relevantes a nível nacional e internacional. Apesar da difusão da disciplina na academia brasileira há mais de meio século, a maioria dos dicionários disponíveis, para uso dos acadêmicos e profissionais da área, são limitados a traduções de obras estrangeiras, em geral, ou obras que objetivam a especialistas, não a leigos. Logo, sem que haja o domínio na área conceitual e terminológica da Linguística, a compreensão desta disciplina é comprometida e até mesmo se torna impossível para os alunos ingressantes em cursos de graduação. No intuito de sanar esta falta de material terminológico bilíngue nesta área, objetivamos desenvolver um vocabulário terminológico bilíngue, português-inglês, na área de Linguística Geral. A metodologia adotada é a de Linguística de Corpus, por meio da qual fazemos a compilação, processamento e análise do *corpus* para extração da lista de palavras, palavras-chave e acesso às linhas de concordância. O console usado para confecção das listas e fases de processamento dos *corpora* é o WordSmith Tools 6.0 (WST 6.0). As teorias que embasam a obra são: (1) a de Barbosa (1990) para o conceito de vocabulário; a de Cabré (1999) com a Teoria Comunicativa da Terminologia (TCT). Barbosa (1990) propõe

que o vocabulário é o léxico de uma língua, cuja unidade seria o vocábulo. Além do vocábulo, concebido como um conjunto, inserido no léxico de uma língua, também existiria o Conjunto Vocabulário ou Vocabulários das Ciências, o vocabulário de especialidades, conjunto terminológico ou terminologia. É dentro dessa perspectiva que optamos por denominar o produto desse projeto Vocabulário da Linguística e não Dicionário de Linguística. No que tange à Terminologia, a TCT de Cabré (1990) propõe que a unidade lexical se caracteriza como termo, devido aos contextos nos quais ela se insere, tornando-se uma unidade terminológica devido a aspectos pragmáticos que caracterizam a comunicação técnico-científica. Ademais, a TCT concebe o termo como uma unidade inserida em um sistema linguístico, ou seja, a unidade terminológica existe dentro de um sistema morfológico, fonológico, semântico e sintático. Os resultados parciais obtidos são os *corpora* de aproximadamente 46 milhões de palavras e manuais de Linguística compilados durante a primeira parte do projeto com discentes da graduação e pós. Neste momento estamos trabalhando na digitalização de manuais de Linguística em português e inglês para completar a quantidade de palavras para a condução final da obra e na padronização dos *corpora* já compilados.

40

FLUÊNCIA E INTERAÇÃO NO INGLÊS PARA AVIAÇÃO: UMA ANÁLISE BASEADA EM PRAGMÁTICA E LINGUÍSTICA DE CORPUS

Malila Prado
Universidade de São Paulo (USP)
malila.prado@gmail.com

O inglês para aviação é um segmento do Inglês para Fins Específicos que merece uma atenção especial devido à sua importância na interação entre profissionais que lidam com segurança de voo. Assim, a Organização de Aviação Civil Internacional (OACI) exige que os governos signatários certifiquem a proficiência linguística em inglês de pilotos e controladores de tráfego aéreo para operações internacionais, respeitando parâmetros propostos em uma escala de proficiência linguística dividida em seis níveis (de 1, elementar, a 6, expert), em seis critérios distintos (pronúncia, estrutura, vocabulário, fluência, compreensão e interação). Na tentativa de compreendermos se e como tais critérios são materializados em comunicações radiotelefônicas autênticas e, assim, buscarmos uma aproximação entre o avaliado / ensinado e o real, propomos uma investigação por meio da correlação entre a Pragmática e a Linguística de Corpus, ou *Corpus Pragmatics* (vide Aijmer e Rühlemann, 2015). A junção dessas linhas de pesquisa possibilitou estudos voltados à identificação dos elementos pragmáticos no contexto específico da comunicação aeronáutica, em um *corpus* com 93 textos transcritos de situações anormais envolvendo a interação entre os profissionais supracitados. Resultados preliminares evidenciaram alta presença de dêiticos (sendo os mais comuns *we*, *you*, *I*, *here*, *now*), o honorífico *sir* e modalizações (*can you*, *would you like*), itens que possibilitam estudos do quesito interação, uma das áreas linguísticas assinaladas pela OACI. O outro campo de interesse desta pesquisa, fluência, foi analisado por meio do levantamento de expressões formulaicas (*that´s*, *yes sir*, *and uh*) na busca por estratégias denominadas fluencemas produtivos

(Götz, 2013). Além disso, será demonstrado como o uso de tais elementos, majoritariamente utilizados quando a troca de informações é sobre o problema em si, funcionam como uma divisão transitória entre o inglês geral e a Fraseologia Aeronáutica. Concluiremos com amostras do *corpus* que justifiquem uma perspectiva mais ampla de avaliação linguística, tal como a competência pragmática, atendendo a princípios intrínsecos desse tipo de comunicação utilizada internacionalmente por uma comunidade profissional.

41

ANÁLISE COMPARATIVA DAS ESCOLHAS LEXICAIS NA TRADUÇÃO DE CONTOS EM LÍNGUA ALEMÃ: UM ESTUDO EXPLORATÓRIO SOB A PERSPECTIVA DA LINGUÍSTICA DE CORPUS

Luciane Leipnitz, Tamara Belmont e Munich Graf
Universidade Federal da Paraíba (UFPB)
luciane.leipnitz@gmail.com

Este trabalho apresenta um estudo exploratório das escolhas lexicais na tradução de um conto da língua alemã para o português brasileiro em dois momentos distintos: o texto traduzido por profissional não tradutor, e o texto final, fruto da revisão por alunos de graduação do Bacharelado em Tradução da UFPB. A revisão do texto traduzido constituiu-se em parceria acadêmica intrainstitucional, independente e voluntária, entre professores e graduandos em Tradução e visou unicamente à produção de um texto final de qualidade na publicação, em língua portuguesa, de um dos Contos de Aldeia de Berthold Auerbach do volume Schwarzwälder Dorfgeschichten (1982). Pretende-se demonstrar que a formação específica em tradução, em oposição a crenças relacionadas ao tradutor e à tradução ainda presentes na visão geral sobre a atividade tradutória (Alves, Magalhães e Pagano, 2003), compreende a aquisição de um conjunto de subcompetências específicas, que constituem a competência tradutória (Pacte, 2003), cujo desenvolvimento depende de orientações teoricamente embasadas e práticas tradutórias reflexivas. O trabalho de revisão, desenvolvido ao longo do período 2016/1, exigiu inicialmente a organização de um *corpus* alinhando texto original e texto traduzido, de modo que as dificuldades encontradas quando da revisão pudessem ser esclarecidas pela consulta ao texto de partida. Com esta breve análise comparativa entre escolhas tradutórias pretende-se mostrar que uma prática tradutória reflexiva começa com um projeto de tradução, ao qual se agregam diversas capacidades e habilidades, desenvolvidas ao longo de uma formação acadêmica específica. Para a análise aqui apresentada, organizamos um *corpus* alinhando texto traduzido-texto revisado, que chamaremos T1 e T2, respectivamente, a partir do qual foi-nos possível realizar buscas com ferramentas online, sistematizando observações levantadas ao longo do processo de revisão. À leitura e marcação de segmentos textuais com alterações substanciais do texto traduzido para a versão final revisada seguiu-se a utilização da ferramenta AntConc, para a geração de listas de palavras e a comparação de escolhas lexicais. As primeiras observações parecem revelar um texto com menor densidade lexical, mais homogêneo quanto a escolhas tradutórias, produzido em português brasileiro padrão, sem marcas de regionalismos, mas ambientado em práticas e costumes alemães de outrora. As análises fazem uso de meto-

dologias da Linguística de Corpus na busca por peculiaridades de textos em língua portuguesa, produzidos por sujeitos distintos, a partir de projetos tradutórios com objetivos diferentes, embora complementares em sua finalidade, por meio do uso de ferramentas de apoio e da adoção de diferentes estratégias para reexpressão na língua de chegada. Para dar continuidade a este estudo exploratório, serão incorporados ao *corpus* outros contos traduzidos a serem ainda revisados, de modo a subsidiar novas pesquisas sobre o processo de aquisição da competência tradutória por meio de formação específica em Tradução.

42

BUSCANDO A REPRESENTATIVIDADE DE UM *CORPUS* ESPECIALIZADO

Anna Maria Maciel, Márcia Moura da Silva e Rozane Rebechi
Universidade Federal do Rio Grande do Sul (UFRGS)
annamb.maciel@gmail.com

Na pesquisa linguística a qualidade de um *corpus* poder espelhar o uso real da língua de dada população é uma questão que até agora parece não ter encontrado consenso. Esta comunicação, gerada em pleno contexto da composição de um *corpus* multilíngue da linguagem da conservação de documentos em suporte papel, pretende discutir se hoje, quando se conta com inimagináveis recursos da informática e progressivos avanços da ciência linguística, ainda é válida a afirmação de que a representatividade do *corpus* é um ato de fé antes que um fato comprovado. Tendo como fundamentação teórica os ensinamentos dos pioneiros que abriram o caminho da atual pesquisa com *corpora* digitalizados, Leech (1991, 2007), Biber (1993), entre outros, apresentamos as linhas mestras que orientam a construção de um *corpus* multilíngue de textos contemporâneos de uma área multidisciplinar técnico-científica. Esse *corpus* deverá ser a principal ferramenta de pesquisa na tarefa de identificação de um repertório terminológico a ser registrado em um glossário online. Nesse contexto, os princípios básicos da operacionalização do design do *corpus* são equacionados com as características do usuário da linguagem a ser representada e com as características dos pesquisadores que a utilizam, levando em consideração, além das especificidades do modo de dizer e os propósitos comunicativos, as condições acadêmicas e logísticas dos responsáveis pelo projeto, sem esquecer desejos e necessidades daqueles que constituem o público preferencial do glossário a ser produzido. A construção do segmento do *corpus* em língua inglesa encontra-se em andamento e é o foco desta comunicação. Tal como a construção dos *corpora* das outras línguas estrangeiras, alemão, espanhol, francês, italiano e russo, foi precedida pelo levantamento da presença de documentos da área na *web*, o que revelou a multiplicidade de ciências e técnicas envolvidas e a diversidade de tipos de texto, de publicações e de comunidades de profissionais interessados. Feito o download, os textos foram examinados e aqueles que satisfizeram os critérios de representatividade à luz dos princípios preestabelecidos foram considerados relevantes e classificados em diferentes categorias antes de serem pré-processados, codificados e armazenados. O conjunto foi então examinado em relação à autoria credenciada dos textos, às condições de produção e recepção, ao balanceamento dos tipos e à relevância da extensão a fim de avaliar o

grau maior ou menor de representatividade da amostragem coletada. Ainda que estejamos em fase de construção do *corpus*, ousamos afirmar que a representatividade de um *corpus* especializado em nossos dias já não é um ato de fé cego, mas um *continuum* que nos permite comprovação e nos faz avançar cada vez mais em direção a seu alcance.

44

O QUE SIGNIFICA(VA) *LOCAL* NO DISCURSO DO MOVIMENTO *FARM-TO-TABLE*?

Stella Tagnin
Universidade de São Paulo (USP)
seotagni@usp.br

O movimento *Farm-to-table* (em português 'do campo à mesa') foi inspirado no livro *The Omnivore's Dilemma*, de Michael Pollan (2006) e propagado por Dan Barber, *chef* americano, com o intuito de valorizar os pequenos agricultores da área onde se localiza seu restaurante, Blue Hill at Stone Barns, em Pocantico Hills, Nova Iorque. Seu objetivo era oferecer a seus clientes produtos os mais frescos possíveis, garantindo, ao mesmo tempo, melhor qualidade do alimento. O movimento teve início ao redor de 2004 (data dos primeiros textos) e estende-se até hoje. Recorrendo à Linguística de Corpus (Baker, 2006; Baker et al., 2008; Tagnin, 2013) pretende-se, a partir de um *corpus* construído com textos em inglês referentes a esse movimento e publicados ao longo desse período, investigar eventuais alterações de sentido que as colocações com o termo *local*, como *local food*, *local produce*, *local farm(s)*, *local farmer(s)*, entre outras, possam ter sofrido. A escolha recaiu sobre *local* por ter sido a palavra-chave que ocorre com o maior número de colocados. A fim de investigar se houve alterações de sentido ao longo dos anos, o *corpus* foi subdividido por períodos: 2004-2010 (primeiros anos), 2011-2014 (auge do movimento e uma nova orientação, apresentada por Dan Barber em 2014, no seu livro *The Third Plate*), 2015-2017 (expansão para fora dos Estados Unidos). A ferramenta usada para a investigação do *corpus* foi o WordSmith Tools, versão 7 (Scott, 2016). Foram analisadas as linhas de concordância das colocações com *local*, expandidas sempre que necessário para melhor esclarecer o sentido. Por vezes, recorreu-se também ao texto todo, conforme sugerem Baker et al. (2008). A análise revelou que alguns restaurantes se apropriaram do termo *local* passando a empregá-lo enganosamente, ou seja, os ingredientes anunciados como 'locais' no cardápio, na realidade não eram obtidos de fontes 'locais'. Atribui-se essa deturpação do sentido a uma apropriação indevida de conceitos de um movimento de sucesso, sem, de fato, fazer uso desses conceitos. Em outras palavras, uma apropriação apenas em benefício (leia-se 'lucro') próprio, contrariando toda a filosofia do movimento *Farm-to-table*.

É POSSÍVEL FALAR EM ESTILO DA TRADUÇÃO EM LEGENDAGEM? UMA PESQUISA BASEADA EM *CORPUS*

Janailton Mick Vitor da Silva e Alessandra Ramos de Oliveira Harden
Universidade de Brasília (UnB)
janailtonm@gmail.com

Este trabalho busca apresentar a pesquisa de mestrado sobre a temática de estilo da tradução no âmbito da legendagem, realizada no Programa de Pós-Graduação em Estudos de Tradução da Universidade de Brasília (POSTRAD/UnB). Filiando-se aos Estudos da Tradução Baseados em Corpus (ETBC), a pesquisa tem, por objetivo, investigar estilo, seja como atributo pessoal ou atributo textual, na tradução de legendas oficiais de filmes. A noção de estilo adotada é a de Baker (2000), que define estilo do tradutor como uma impressão digital, descrevendo seu comportamento linguístico a partir de padrões recorrentes, estatisticamente observáveis, que ilustram uma preferência consistente por itens lexicais específicos, padrões sintáticos, recursos coesivos ou estilo de pontuação, quando outras opções estariam igualmente disponíveis na língua. Além disso, pretende-se construir um arcabouço teórico que dialogue pesquisas em estilo da tradução e do tradutor literário (Baker, 1999, 2000, 2004, 2007; Leech e Short, 2007; Malmkjær, 2003, 2004; Saldanha, 2011a, 2011b) com a tradução para legendagem oficial (Bogucki, 2009; Díaz Cintas, 2005a; Díaz Cintas e Remael, 2007; Gottlieb, 1994, 1998, 2005a, 2005b; Ivarsson e Carroll, 1998). Será construído um *corpus* paralelo (Baker, 1995), composto por falas em inglês e suas respectivas traduções para o português brasileiro, material esse extraído de cinco filmes do mesmo gênero e legendados pelo(a) mesmo(a) tradutor(a) profissional nos últimos cinco anos. Para auxiliar na investigação de traços que possam indicar estilo, serão utilizados três utilitários do WordSmith Tools, versão 7.0 (Scott, 2017): i) Viewer & Aligner: alinhamento de textos fontes com textos alvos; ii) WordList: compilação de listas de palavras para obter dados acerca da variedade lexical (razão forma/item e razão forma/item padronizada), tamanho dos textos, número de sentenças, entre outros; iii) Concord: visualização de palavras e frases acompanhadas do contexto em que ocorrem, a partir de busca por nódulo (palavras e etiquetas). Apesar das limitações metodológicas que o estudo de estilo enfrenta no âmbito da tradução, como apontam as pesquisas nessa área, busca-se verificar a ocorrência de padrões estilísticos consistentes e se esses padrões se verificam em mais de uma tradução do(a) mesmo(a) legendista, tentando, ainda, detectar se esses padrões podem ser atribuídos a(o) tradutor(a), levando à elaboração de um perfil estilístico individual, ou às traduções. Intenciona-se, por fim, contribuir para uma epistemologia de campo, suscitando questionamentos acerca da possibilidade do estudo de estilo em legendagem, como atributo pessoal ou textual.

UTILIZANDO *CORPORA* NA PESQUISA DA TERMINOLOGIA DE GÊNERO NA LEGISLAÇÃO BRASILEIRA

Clarissa Isabel V. de Oliveira
Universidade Federal do Rio Grande do Sul (UFRGS)
clarissaisabel@gmail.com

Este trabalho em andamento se insere na pesquisa do reconhecimento da terminologia referente às questões de gênero nos textos legislativos brasileiros de âmbito federal e do Rio Grande do Sul. Definimos essa terminologia como itens lexicais relacionados às mulheres e às pessoas LGBT, bem como aos contextos sociais e jurídicos que motivam a produção legislativa voltada a esses sujeitos de direito. Considerando a Terminologia como janela para compreensão do mundo, objetivamos, em nossa apresentação, demonstrar que em *corpora* de textos legislativos encontramos registros terminológicos de transformações sociais. Ponto central de nosso estudo é o termo de natureza categorial "gênero". Partimos do suposto de que ao redor desse termo gravitam os termos distribuídos em áreas temáticas, tais como trabalho, violência e direitos reprodutivos. Nossa abordagem relaciona-se com a Semântica de *Frames*, segundo a qual o processo de pareamento de uma palavra com seu *frame* subjacente requer que se reconheça a relevância da informação subjacente, dentro da qual essa palavra desempenha um papel interpretativo (Fillmore e Baker, 2009, p. 317-318). Em nosso estudo, essa perspectiva dialoga com a Teoria Sociocognitiva da Terminologia (TST), que trata os termos como unidades de compreensão (u.c.), as quais funcionam em modelos cognitivos e evoluem constantemente por questões sócio-históricas (Temmerman, 2004, p. 35), correspondendo ao caráter sociológico de nosso objeto. Metodologicamente, utilizamos *corpora* numa abordagem interdiscursiva. Constituímos um *corpus* de estudo (*corpus* A) composto de 90 textos legislativos relacionados às questões de gênero, e, como apoio para compreender o conhecimento subjacente às u.c. observadas no *corpus* A, um *corpus* subsidiário (*corpus* B), composto de textos de referência, de popularização e científicos sobre questões de gênero, os quais assumimos como representativos da área e como influenciadores de conteúdo dos textos legislativos. Para compor o *corpus* B, consideramos, conforme Sinclair (2004), que textos de teóricos reconhecidos ou de fontes oficiais constituem critério de representação externa. Identificamos no *corpus* de textos legislativos candidatos a u.c., cujos critérios de seleção são, até o momento: termos que façam referência aos sujeitos de direitos (e.g., vítimas de violência sexual, que designa o sujeito de direito mulher), aos seus direitos (e.g., espaço de escuta) e a fatores motivadores de direitos (e.g., feminicídio). No *corpus* B, com auxílio do software AntConc, observamos concordâncias para o termo "gênero" como nódulo de busca dos elementos linguísticos que integram a estrutura do *frame* GÊNERO. O uso de *corpora* nos possibilita, como resultados preliminares: 1) constatar a ocorrência de variação semântica nos textos legislativos – a uma mesma forma lexical correspondem uma u.c. de natureza categorial e uma u.c. de natureza conceitual com função identitária; 2) identificar a estrutura do *frame* GÊNERO, cujos elementos são, entre outros, mulheres, pessoas LGBT, homens e Estado, no papel de participantes, e trabalho, violência e discriminação no papel de acessórios; e, além disso, 3) demonstrar a

importância comunicativa (Sinclair, 2004) do *corpus* subsidiário na busca da plena compreensão das u.c., dada a ausência de definição do termo "gênero" no *corpus* A.

57

A MODALIZAÇÃO AVALIATIVA EM RESUMOS CIENTÍFICOS: UM ESTUDO DA LÍNGUA ESPANHOLA COM BASE EM *CORPUS*

Mauren Thiemy Ito Cereser e Sandra Dias Loguercio
Universidade Federal do Rio Grande do Sul (UFRGS)
mauren.cereser@gmail.com, sandraloguercio@hotmail.com

O presente estudo faz parte de um trabalho maior que analisa e descreve o funcionamento dos modalizadores de maneira contrastiva em textos científicos em língua portuguesa, espanhola e francesa para fins pedagógicos e/ou de auxílio à tradução. A modalização pode ser compreendida como uma estratégia argumentativa em que o locutor assume um posicionamento perante o enunciado, imprimindo uma avaliação ou ponto de vista sobre o que é dito (Nascimento e Lima, 2012). Esse posicionamento, segundo Castilho (2010) e Nascimento e Lima (2012), pode ser expresso de três maneiras: i. pela modalização epistêmica (em que o conteúdo proposicional é avaliado em relação a seu valor de verdade); ii. deôntica (em que o conteúdo proposicional é entendido como um estado de coisas obrigatório, necessário ou possível); iii. avaliativa ou afetiva (em que há um julgamento de valor em relação ao conteúdo proposicional). Textualmente, os modalizadores são marcados lexical e/ou morfologicamente, pelo uso de incisas, por marcas tipográficas ou prosódicas (Kerbrat-Orecchioni, 1980); podem aparecer de modo mais explícito (como nas modalidades frasais, em verbos que denotam uma operação lógica, etc.) ou estarem incorporados ao dito (emprego de certos tempos verbais, uso de adjetivos ou advérbios, por exemplo). Para este estudo, analisaram-se resumos da área de Linguística a partir de *corpora* em português e espanhol compostos por 100 textos, enfocando-se especificamente a modalização avaliativa, expressa sobretudo por adjetivos e/ou advérbios. Esse tipo de modalização, por evidenciar julgamentos de valor e, desse modo, revelar um sentimento do locutor, é o menos provável no discurso científico ou aquele que se busca evitar, em nome do rigor e da busca pela imparcialidade da ciência. Porém, qualificações como "importante", "relevante", "valioso", entre outras, também integram tal discurso, que é construído de modo a obter a adesão do leitor. A análise proposta se divide em três etapas: i. familiarização com os resumos (estudo da macroestrutura e identificação de índices de modalização); ii. extração automatizada de informação linguística com uso do programa AntConc (versão 3.4.4, especialmente das ferramentas Concordance e Wordlist); iii. análise e contraste dos modalizadores avaliativos frequentes nas línguas abordadas. Os resultados preliminares indicam que esse tipo de modalização ocorre mais frequentemente em resumos produzidos em língua espanhola, sendo mais raros em resumos escritos em língua portuguesa.

O PROCESSO DE ORGANIZAÇÃO E ETIQUETAGEM DO *CORPUS* LINGUÍSTICO "UFSM"

Sabrina Damiani
Universidade Federal de Santa Maria (UFSM)
sabrinadamiani.12@hotmail.com

Com a finalidade de responder às demandas em pesquisas nos níveis de graduação e pós-graduação dos cursos de Letras da UFSM, em especial às investigações desenvolvidas na área de Linguística Sistêmico-Funcional (Halliday, 1994, 2004; Halliday e Matthiessen, 1999, 2004, 2014), especialmente em língua portuguesa, procedemos, com o auxílio da ferramenta computacional WordSmith Tools (Scott, 2009), à coleta e organização de um *corpus* que denominamos "*Corpus* Linguístico UFSM". Nosso objetivo principal é compor um *corpus* de textos de cunho acadêmico, jornalístico e político, que são os mais investigados pela linha de pesquisa Linguagem no Contexto Social à qual este trabalho está vinculado. Os textos de cunho jornalístico e político foram escolhidos por serem dois gêneros estudados por Martin e Rose (2008), já os textos de acadêmicos foram escolhidos pelos pesquisadores de Linguagem no Contexto Social, pois trata-se de um gênero ainda pouco explorado na LSF. Inicialmente, realizamos a coleta dos textos em *sites* da mídia (jornais do centro do país e também do Rio Grande do Sul) e de instituições políticas (*site* da Presidência da República, Biblioteca do Congresso Nacional, páginas específicas de representantes políticos no Brasil) e universitária (especialmente de cursos de pós-graduação), os quais foram devidamente separados em três blocos (acadêmico, jornalístico e político). Posteriormente, para cada bloco foram criadas subpastas, seguindo o critério da divisão por gêneros textuais (Martin e Rose, 2008, Rose e Martin, 2012). Até o momento coletamos 500 textos que constituem o *corpus* formado por mais de 1.000.000 de *tokens*. Em um terceiro momento, passamos à fase de etiquetagem, o que estamos realizando com o auxílio do etiquetador online VISL (Visual Interactive Syntax Learning), desenvolvido por pesquisadores dinamarqueses (Sardinha, 2004). Como o *corpus* precisa ser constantemente alimentando, traçamos como meta, até o final de 2017, reunir um total de 800 textos devidamente etiquetados, de modo a prover suporte aos trabalhos desenvolvidos em análises de discurso pelos pesquisadores dos cursos de Letras da Universidade Federal de Santa Maria. Também é nosso objetivo socializar nossas investigações, de modo a tornar esta coletânea disponível na internet, a fim de que seja utilizada por usuários interessados no trabalho com textos das três áreas apontadas.

ESPECIFICIDADES DO VOCABULÁRIO DA OBRA *CONTOS GAUCHESCOS*, DE SIMÕES LOPES NETO: UM ESTUDO EXPLORATÓRIO

Nathalia Marcon e Maria José Finatto
Universidade Federal do Rio Grande do Sul (UFRGS)
ntolivam@gmail.com, mariafinatto@gmail.com

INTRODUÇÃO: O objetivo deste trabalho é apresentar uma análise inicial, com apoio de recursos da Linguística de Corpus, do vocabulário empregado na obra literária *Contos Gauchescos* (1912), de Simões Lopes Neto (1865-1916), importante autor da literatura sul-riograndense, visto que o acesso desse livro tende a ser considerado como bastante difícil para leitores atuais (conforme Fischer, 2000, p. 7-8). Ao descrever o vocabulário da obra, apontam-se possíveis especificidades do gênero literário e do autor frente a autores seus contemporâneos (em um *corpus* de contraste) e frente a dois *corpora* gerais, não literários, um do mesmo período da produção da obra e um *corpus* atual do português brasileiro (*corpora* de referência). MATERIAIS E MÉTODOS: A observação do vocabulário empregou a ferramenta AntConc 3.4.1m. mediante os recursos lista de palavras, relação *types*/*tokens* e lista de palavras-chave. O livro *Contos Gauchescos* apresenta 32.879 *tokens* e 6.428 *types*; e teve seu vocabulário contratastado com aqueles das seguintes obras: *Triste Fim de Policarpo Quaresma* (1915): 66.683 *tokens* e 10.713 *types*, *Os Sertões* (1902): 154.735 *tokens* e 22.134 *types* e *Iracema* (1865): 29.898 *tokens* e 5.563 *types*; como *corpora* de referência utilizamos: (a) cartas de diferentes remetentes escritas entre os anos de 1800 e 18995, disponíveis no Corpus Histórico do Português Tycho Brahe; (b) listas de palavras (da escrita e da fala) do Banco de Português. Inicialmente, foram feitas listas de palavras dos *corpora* de estudo e de contraste. Em seguida, foi criada uma *keyword list*, contrastando esses dois *corpora*. Dessa lista, as primeiras 1750 palavras (31%) foram comparadas com os *corpora* de contraste e de referência, identificando-se os itens de frequência zero em ambos os *corpora*. RESULTADOS PARCIAIS: Os resultados da tabela 1 sugerem que o texto dos *Contos Gauchescos* não possui uma grande riqueza lexical, mas é composto por vocábulos específicos: os 5.647 vocábulos da *keyword list*, que representam as palavras cuja frequência é estatisticamente diferente entre os *corpora* de estudo e de contraste, correspondem a 87,8% dos *types* do *corpus* de estudo. Essa idéia é reforçada pelos resultados da tabela 2, que destaca a quantidade de vocábulos de frequência zero na comparação entre os *corpora*. A análise realizada não pretendeu esgotar a investigação sobre a especificidade lexical encontrada, merecendo o presente trabalho um maior aprofundamento. Identificamos, por exemplo, palavras como 'china', que, embora conste na lista de palavras do *corpus* de referência, é plausível supor-se que com significado diverso daquele empregado por S. Lopes Neto, a saber: mulher, especialmente a mestiça ou a índia. Assim, a possibilidade de lematização e análise semântica dos vocábulos deve ser levada em consideração em uma análise futura.

PESQUISA LINGUÍSTICA COM *CORPORA* E MEMÓRIAS DE TRADUÇÃO

Liana Paraguassu
Universidade Federal do Rio Grande do Sul (UFRGS)
liana@linguatraducoes.com

Este trabalho visa a demonstrar de que forma as memórias de tradução (MTs), gerenciadas pelas CAT Tools (Computer Assisted Translation Tools), podem ajudar o tradutor que adota a abordagem metodológica da pesquisa linguística com *corpus* e os pressupostos fundamentais dos Estudos de Tradução e da Terminologia, a realizar um trabalho mais eficaz. As MTs são uma base de dados que armazena os dados do texto fonte a ser traduzido e do texto alvo já traduzido, nos pares de idiomas de trabalho, em segmentos paralelamente alinhados, formando, assim, um *corpus* paralelo (Kawamorita e Caseli, 2012; Nogueira, 2012). As CAT Tools trabalham com um algoritmo de correspondência; este algoritmo é capaz de reconhecer se os segmentos de um texto a ser traduzido correspondem totalmente (100% *match*) ou parcialmente (*fuzzy match*) aos dados de textos já traduzidos e armazenados anteriormente na MT. O principal objetivo das MTs e das CATs é aumentar o rendimento do tradutor por meio da reciclagem de textos já traduzidos, gerando, com isso, uma significativa economia para as empresas que contratam serviços de tradução. Contudo, essas ferramentas, sozinhas, não são capazes de auxiliar o tradutor na busca por equivalentes tradutórios, por isso a importância de o tradutor também utilizar outros *corpora* para aprimorar seu trabalho e procurar subsídios para um equivalente tradutório mais adequado. Sendo assim, concluímos que é preciso que o tradutor saiba desenvolver, juntamente com a competência tradutória nos pares de idioma de trabalho, mais uma relativa competência informática e o conhecimento teórico e aplicado da pesquisa linguística com *corpus* (Frerot, 2016; Garcia, 2015). Este trabalho em andamento tem por objetivo demonstrar de que forma as memórias de tradução e as CAT Tools podem trabalhar em conjunto com ferramentas de linguística de *corpus* para ajudar o tradutor nesse processo. Para tanto, exemplificaremos como utilizamos um *corpus* de textos de mecânica, especificamente de motores, armazenado na memória de tradução da ferramenta SDL Trados Studio 2017 para decidir sobre um equivalente tradutório. O primeiro passo foi aplicar as diferentes ferramentas do software AntConc no *corpus* paralelo do Trados Studio para definir quais palavras mais se destacavam. Em seguida, selecionamos *crankshaft* como termo de trabalho, por este apresentar diferentes traduções (virabrequim, girabrequim, eixo de manivelas) para o mesmo termo em inglês e para designar a mesma peça do motor. Após, utilizamos a ferramenta *Sketch Engine* para analisar as ocorrências desses termos em um *corpus* de referência em língua portuguesa e compará-las com nossos dados. A verificação de sua frequência juntamente com o padrão de comportamento lexical nos fez optar por virabrequim como o equivalente tradutório mais adequado. Com isso, concluímos a importância da pesquisa com *corpora* pelo tradutor, em paralelo ao uso das MTs, de modo que ele possa definir qual termo deve ser utilizado como tradução na língua de chegada, uma vez que o universo das memórias de tradução é muito menos representativo do uso real de um termo do que aquele dos *corpora* de grandes proporções encontrados na *web*.

ANALISANDO PACOTES LEXICAIS EM UM *CORPUS* MULTINACIONAL DE PORTUGUÊS ACADÊMICO

Tanara Zingano Kuhn, José Pedro Ferreira, Margarita Correia
CELGA-ILTEC - Centro de Estudos de Linguística Geral e Aplicada
da Universidade de Coimbra, Portugal. Bolsista da Capes-Brasil
Faculdade de Letras da Universidade de Lisboa, Portugal
tanarazingano@outlook.com, jpf@uc.pt, margarita@campus.ul.pt

Vários estudos demonstram que os pacotes lexicais (*lexical bundles*), entendidos como sequências recorrentes de palavras, normalmente incompletas e sem caráter idiomático, variam conforme o registro em que são empregados (Biber *et al.*, 1999; Biber, 2009). Assim, quando usados em textos acadêmicos, os pacotes apresentam diferentes formas, padrões, frequências e funções que permitem, de um lado, distinguir esse registro de outros (Biber *et al.*, 1999; Biber, 2006), e, de outro, contribuir para a diferenciação entre gêneros (Hyland, 2008), disciplinas (Biber e Barbieri, 2007) e modo oral ou escrito (Biber, 2009). Embora a grande maioria dos estudos sobre os usos de pacotes lexicais no âmbito acadêmico diga respeito à língua inglesa, revisão bibliográfica indica que esse tópico tem sido alvo de interesse crescente também no que tange ao português acadêmico (Kilian e Loguercio, 2015). Nesse sentido, o trabalho em desenvolvimento que aqui se apresenta pretende contribuir para esse campo através de um estudo exploratório acerca do uso de pacotes lexicais (cf. Biber *et al.*, 2004) no Corpus de Português Escrito em Periódicos (CoPEP) (Kuhn e Ferreira, 2016), que contém cerca de 40 milhões de palavras, abrange seis grandes áreas do conhecimento e é equilibrado entre as variedades português brasileiro (PB) e europeu (PE). Baseando-nos nos estudos de Biber *et al.* (2004 e 2007) e Sardinha *et al.* (2015), propomos uma pesquisa dividida em três etapas. A primeira tem como objetivos identificar os pacotes lexicais formados por quatro palavras que ocorrem no CoPEP, observar seus padrões estruturais de forma a categorizá-los em diferentes tipos formais e apontar sua distribuição em termos de áreas de conhecimento e variedades PB e PE. Já a segunda fase diz respeito ao desenvolvimento de uma taxonomia das funções discursivas dos pacotes lexicais encontrados no CoPEP e descrição de sua distribuição no *corpus*. Por fim, a terceira etapa consiste em analisar e discutir os resultados, tendo em vista os aspectos formais e funcionais dos pacotes lexicais, verificando se estes contribuem para a caracterização de textos acadêmicos escritos em PE e PB e no escopo das diferentes áreas de conhecimento. Pretendemos enriquecer essa discussão ao cotejarmos nossos resultados com aqueles obtidos em outros estudos com o mesmo objeto, salvaguardando, evidentemente, as diferenças de âmbito metodológico. Entre estes, destacamos aqui a investigação desenvolvida por Sardinha *et al.* (2015, p. 56), que, adotando a taxonomia de funções dos pacotes lexicais proposta por Biber *et al.* (2006), apontou a expressão de referência como sendo a função mais frequente dos pacotes lexicais em artigos científicos escritos em PB. Outro importante estudo a ser usado para fins comparativos é o de Bennett (2010), no qual a autora constata que textos acadêmicos escritos em PE, ao contrário daqueles escritos em inglês, tendem a apresentar observações e afirmações de forma indireta, subordinadas a orações principais que

enfatizam a dimensão interpessoal (Bennett, 2010, p. 25), normalmente formando estruturas excessivamente elaboradas (*e.g.*, "não deixa de ser interessante notar que"). Com o intuito de contribuir para os estudos acerca do português acadêmico, traremos aqui alguns resultados, ainda que preliminares, de nossa pesquisa. Apoio: **CAPES**

69

O USO DE *CORPUS* NO DESENVOLVIMENTO DE DICIONÁRIOS ELETRÔNICOS ORGANIZADOS A PARTIR DA NOÇÃO DE *FRAME*

Bruna da Silva
Universidade do Vale do Rio dos Sinos (Unisinos)
broonamoraes@gmail.com

A Semântica de *Frames* é uma teoria de base empírica que defende que somente é possível compreender o significado de uma palavra ao relacioná-la a um sistema estruturado de conhecimento, ou seja, a um *frame*. Este trabalho, que se insere na interface entre Semântica de *Frames* (Fillmore, 1982; 1985) e Lexicografia Eletrônica, tem como objetivo investigar a contribuição do uso de *corpus* no desenvolvimento de dicionários eletrônicos organizados a partir da noção de *frame*, visto que tanto a Semântica de *Frames* quanto a Lexicografia preveem em suas metodologias de trabalho a utilização de dados provenientes de *corpus* como forma de acesso a evidências empíricas sobre os usos linguísticos. A metodologia inclui a análise do Dicionário Olímpico (Chishman, 2016), produto desenvolvido pelo grupo de pesquisa SemanTec, da Unisinos, que descreve os *frames* (cenários) de cada uma das modalidades olímpicas e apresenta suas respectivas unidades lexicais (palavras). As unidades lexicais do dicionário, assim como seus equivalentes de tradução para a língua inglesa e os exemplos em língua inglesa provêm de *corpora* compilados para servir a tal finalidade. Os objetivos metodológicos incluem identificar em que medida os *corpora* foram utilizados e para quais finalidades seu uso estava voltado. Cumpridos os objetivos, pretende-se avaliar de que modo esses resultados contribuíram para o produto final. Embora a relação entre lexicografia impressa e uso de *corpus* não seja algo novo (considerando-se os muitos aspectos que esse emprego atende na criação de um dicionário tradicional, tais como fornecer lista de palavras, equivalentes de tradução, exemplos, etc.), a Lexicografia Eletrônica (nesse caso, aliada à Semântica de *Frames*) abre uma série de possibilidades para o uso de evidências provenientes de *corpus* no desenvolvimento de dicionários. As mais importantes dessas possibilidades se relacionam a (i) mostrar conexões entre significados e padrões de uso e (ii) identificar os contextos nos quais as palavras são utilizadas. O trabalho está em fase de desenvolvimento. Palavras-chave: Semântica de *Frames*. Lexicografia Eletrônica. Linguística de Corpus.

75

A CORPUS ANALYSIS OF THE VOCABULARY FROM BEYONCÉ'S LEMONADE

Natasha Barth Sidoruk e Aline Xavier de Vargas
Universidade Federal do Rio Grande do Sul (UFRGS)
natasha.sidoruk@hotmail.com, alinexv00@gmail.com

Lemonade (2016) is Beyoncé's latest album. It is different from her previous work in that it addresses themes that had not been encompassed by the artist before, such as abusive relationships and racial and gender inequality. In order to determine if the change in theme affected the vocabulary used in the lyrics, two corpora *were compiled: (i) the reference corpus, which corresponded to her previous albums, and (ii) the study corpus, which contained Lemonade's song lyrics. The keyword list revealed significant differences between the two* corpora*. Four vocabulary categories were set out in Lemonade's keyword list, related to the central subjects of the album.*

76

QUANDO A GLOBALIZAÇÃO AFLIGE: UM ESTUDO CONTEMPORÂNEO DAS VARIEDADES FONÉTICO-FONOLÓGICAS EM LÍNGUA INGLESA BASEADO EM *CORPUS*

Adeilson de Jesus Oliveira Júnior
Universidade do Estado da Bahia – Campus XIV. Conceição do Coité, BA
adeilson.junior@hotmail.com.br

The purpose of this paper is to investigate through analysis based on Corpus Linguistics, some differences that arise in phonetics of the English Language. We aimed to present a brief theoretical review of the English language history and its "status", considering the changes that occur mainly with the expansion of the English language in globalizing processes. It is our aim to consider the analyzes compiled from online *Corpora* such as VOICE, the Vienna University, coordinated by the researcher Barbara Seidlhofer.

81

ARTIGOS CIENTÍFICOS SOBRE TRANSTORNO DO ESTRESSE PÓS-TRAUMÁTICO: DESAFIOS DA ACESSIBILIDADE LINGUÍSTICA PARA PÚBLICO LEIGO

Asafe Davi Cortina Silva
Programa de Pós-Graduação do Departamento de Letras da Universidade Federal do Rio Grande do Sul (UFRGS)
asafedcs@gmail.com

O presente trabalho descreve o estudo-piloto de um projeto de dissertação cujo principal objetivo é analisar a complexidade de artigos científicos sobre o Transtorno do

Estresse Pós-Traumático, tanto em português como em inglês, para investigar características textuais que possam aumentar ou reduzir a complexidade de um texto. Uma vez que esse projeto se encontra em seus passos iniciais, ele descreve apenas a análise de algumas métricas específicas e suas possíveis indicações.

84

ATIVIDADES DE COMPREENSÃO ORAL COM BASE EM *CORPORA*: UM ESTUDO PILOTO

Luciano F. da Silva, Paula T. Pinto e Elen Dias
Universidade Estadual Júlio de Mesquita Filho; Universidade Estadual Júlio de Mesquita Filho (UNESP – IBILCE); Faculdade de Tecnologia de Jales (FATEC)
luciano.francco@gmail.com, paulapintounesp@gmail.com, elen.dias@fatec.sp.gov.br

O presente artigo é um recorte de uma pesquisa de mestrado que tem como objetivo a criação de atividades com base em *corpora*, voltadas ao desenvolvimento da compreensão oral para alunos em nível A2 e B1. Para tanto, foram utilizados os conceitos da Linguística de Corpus (O'Keeffe, McCarthy e Carter, 2007; Berber Sardinha, 2004; Sinclair, 2004), além de Inglês para Fins Acadêmicos (Charles, 2013; Flowerdew, 2001) e Desenvolvimento da Compreensão Oral em Língua Inglesa (Goh, 2012; Buck, 2011; Rost, 2011). Como embasamento teórico-metodológico para elaborar as atividades que serão apresentadas em um minicurso ministrado em uma instituição de ensino superior no noroeste do estado de São Paulo. As atividades foram criadas com base em dois *corpora* compilados a partir de apresentações orais do TED Talks e TED-Ed.

85

O USO DE *CORPORA* MULTIMODAIS NA CONSTRUÇÃO DE UM DICIONÁRIO BASEADO NA NOÇÃO DE *FRAMES* SEMÂNTICOS

Sandra de Oliveira
Universidade do Vale do Rio dos Sinos (Unisinos)
sandra_san05@hotmail.com

Vinculado ao projeto Dicionário Modalidades Olímpicas 2016 (Chishman, 2015), cujo propósito foi criar um dicionário de termos olímpicos em português e inglês baseado na noção de *frames* semânticos (Fillmore, 1982), este estudo tem como tema o processo de coleta dos *corpora* em português que forneceram os dados para a construção dos *frames* relativos aos esportes olímpicos. Justifica-se pelo fato de fornecer dados que contribuem para a construção de um dicionário que auxilie as pessoas no processo de construção dos sentidos das palavras desses domínios lexicais. O objetivo deste trabalho é discutir como se deu esse processo de coleta dos *corpora*: seleção e armazenamento dos textos. A metodologia se apoiou na teoria semântica de Fillmore e fez uso do software AntConc como recurso de exploração dos *corpora*. Consistiu na coleta de textos do gênero *match reports*, que relatam as ações desempenhadas dentro dos esportes. Foi efetuada a seleção de *sites* e registrados em uma tabela de *links* para otimizar o mo-

mento da coleta. A partir da tabela, as notícias de cada *site* foram examinadas e as que atendiam aos critérios foram selecionadas. O modo de armazenamento consistiu em salvar cada texto individualmente em formato .txt, codificação UTF-8, ignorando informações que não faziam parte do relato, como imagens e comentários. Os resultados apontam para o fato de que os critérios de seleção dos textos não puderam ser seguidos à risca, uma vez que fazem parte das modalidades olímpicas esportes que não possuem nenhum tipo de narração, seja por consistirem em ações simplificadas, como partida e chegada, ou por não serem difundidos no Brasil. Assim, a adoção dos gêneros multimodais foi necessária, como foi o caso do Rugby 7s e de outros esportes. Recorreu-se a manuais de regras oficiais, textos didáticos disponibilizados por sites das confederações e por profissionais da área. A leitura desses textos, além da coleta das unidades lexicais de cada modalidade, possibilitou a compreensão da dinâmica dos esportes. Isso permitiu a construção de mapas conceituais que organizaram as informações enciclopédicas de cada esporte e facilitaram a identificação dos seus eventos (*frames*). Como complemento, foram assistidas partidas do Campeonato Mundial e do Sulamericano narradas em português, tanto no YouTube como em canais da TV aberta, nos quais, ao serem identificados os *frames*, foram transcritas as sentenças que evidenciam os elementos pertencentes a esse *frame* e seus relacionamentos. Enfim, a partir desses *corpora*, através do software AntConc, foram extraídas todas as unidades lexicais e as informações que possibilitaram a definição dos cenários do Dicionário Olímpico. As estratégias utilizadas possibilitaram que todas as informações presentes no Dicionário Olímpico tenham sido obtidas de textos extraídos do mundo real, o que vem ao encontro do posicionamento da Semântica de *Frames* e, de forma geral, da Linguística Cognitiva.

87

O USO DE *CORPUS* PARALELO E COMPARÁVEL PARA DESCREVER PADRÕES DE USO NA TRADUÇÃO DE ABREVIATURAS E ACRÔNIMOS DE TERMOS MÉDICOS

Márcia Moura da Silva, Gabriele Paparelli e Gislaine Alves
Universidade Federal do Rio Grande do Sul (UFRGS)
marciamouras@hotmail.com

Basta uma busca pelas dissertações e teses defendidas nos programas de pós-graduação das universidades brasileiras para se constatar que, embora as pesquisas com tradução de textos médicos tenham aumentado nos últimos anos, ainda há pouco interesse acadêmico em investigar as escolhas tradutórias para abreviaturas e acrônimos, elementos recorrentes nesses textos. Ainda que haja, como nos lembra Aixelá (2009), certa disposição em manter abreviaturas inalteradas em textos traduzidos do inglês para manter o caráter de dada disciplina e facilitar a divulgação do conhecimento técnico-científico, a prática nos mostra que esses elementos estão longe de serem estanques e que tampouco há consenso em relação a como traduzi-los, o que acaba por dificultar o trabalho do tradutor. Como grande parte da divulgação do conhecimento técnico-científico é feita em língua inglesa, a tradução de artigos médicos para essa língua vem se tornando imprescindível. Assim sendo, o objetivo desta pesquisa é verificar o comportamento tradu-

tório em relação a abreviaturas e acrônimos no par linguístico português-inglês e propôr um glossário online que sirva como fonte de pesquisa para tradutores em formação, tradutores profissionais, redatores e revisores. Além da tradução técnico-científica (Azenha, 1996 e 1999; Aixelá, 2009; Aubert, 2001), a pesquisa tem como base teórica os Estudos da Tradução Baseados em Corpus (Baker, 1993 e 1995; Olohan, 2004; Tymoczko, 1998), que emprestam da Linguística de Corpus (Baker, 2013; Berber Sardinha, 2002 e 2004; Leech, 1991; Tognini-Bonelli, 2001) alguns de seus princípios e algumas das técnicas por ela desenvolvidas, sobretudo no que diz respeito à metodologia de compilação de *corpus*. A pesquisa se apoiará na compilação de três *corpora*: i) *corpus* paralelo com textos originais em língua portuguesa e suas traduções em língua inglesa da Revista Brasileira de Reumatologia (RBR); ii) *corpus* comparável monolíngue com os textos traduzidos em língua inglesa da RBR, que serão comparados aos textos escritos originalmente em inglês da revista Rheumatology; e iii) *corpus* comparável bilíngue, composto de textos escritos originalmente em inglês da Rheumatology e textos escritos originalmente em português da RBR. A combinação desses *corpora* possibilitará verificar se determinadas abreviaturas/acrônimos são mantidas inalteradas ou traduzidas e os padrões de uso dos mesmos, indicando se determinado padrão está restrito ao texto traduzido ou se é mais frequente nele. Espera-se que tal conhecimento possa auxiliar tradutores e pesquisadores na produção de textos que tenham mais aceitação na comunidade internacional. Para Olohan (2004), a influência da língua de partida em padrões de uso observados na língua de chegada seria um dos principais benefícios da combinação entre esses dois tipos de *corpora*. Porém, visto que as traduções objeto desta pesquisa são em língua inglesa, e dada sua influência na área aqui pesquisada, espera-se o cenário inverso, ou seja, maior influência da língua inglesa sobre as abreviaturas e acrônimos usados em língua portuguesa, embora já se tenha observado um número considerável de abreviaturas e acrônimos em português, como é o caso de AINEs (anti-inflamatórios não esteroides) e AVAQ (ano de vida ajustado pela qualidade) para NSAIDs (non-steroidal anti-inflammatory drugs) e QALYs (quality-adjusted life years), respectivamente.

89

SOBRE A COMPILAÇÃO DE UM *CORPUS* PORTUGUÊS-INGLÊS DE RESENHAS DE RESTAURANTES PARA FINS TRADUTÓRIOS

Rozane Rebechi, Ana Cristina Neves Nachtigall,
Hannah dos Santos Kahn e Nathália Glasenapp
Universidade Federal do Rio Grande do Sul (UFRGS)
nachtigall.ana@gmail.com

Tema recorrente em programas de TV, blogs, livros e revistas, o texto culinário não desfruta de semelhante *status* em pesquisas acadêmicas, possivelmente pelo fato de esse gênero textual ter sido por muito tempo considerado inferior, se comparado à literatura e a outras artes (Cf. Capatti e Montanari, 1999). Contudo, com a globalização, textos culinários são constantemente traduzidos de uma língua para outra, evidenciando falta de padronização na escolha dos equivalentes de referências culturais, ausência de naturalidade, definições inconsistentes, descaracterização de marcas culturais e até mesmo erros gros-

seiros, que, obviamente, prejudicam a compreensão por parte do leitor (Rebechi, 2015). A pouca oferta de materiais de referência para tradutores e redatores da área agrava os problemas. Como parte do projeto "Culinária para Fins Acadêmicos: compilação de um *corpus* de textos culinários com foco na tradução", desenvolvido junto ao Instituto de Letras da UFRGS, esta pesquisa em andamento, que visa a abordar a compilação e a análise preliminar de um *corpus* comparável bilíngue português-inglês de resenhas de restaurantes para a extração de termos em português, e equivalentes, fraseologias e exemplos de uso em inglês, apoia-se na metodologia subjacente à Linguística de Corpus (Sinclair, 1991; Bowker e Pearson, 2002), na abordagem funcionalista da tradução (Nord, 1991) e na Teoria Comunicativa da Terminologia (Cabré, 2005). Para atingir os objetivos propostos, foram selecionadas resenhas publicadas em periódicos brasileiros e norte-americanos de grande circulação. Cada *subcorpus* contabiliza aproximadamente 300.000 palavras (*tokens*) e foi processado por meio da ferramenta WordSmith 6.0 (Scott, 2012). O levantamento das palavras-chave em português mostrou que dois termos são utilizados para se referir ao estabelecimento: 'casa' e 'restaurante'. Já em inglês, utiliza-se apenas restaurant para o mesmo fim. Nas resenhas em português, é recorrente a expressão 'acesso para deficientes'. Já no *subcorpus* em inglês, foi identificada recorrência de *wheelchair access*, equivalente que se mostra apropriado para garantir a convencionalidade da tradução ou redação em língua inglesa. O levantamento de combinações de palavras (*clusters*) também permite que se identifiquem fraseologias equivalentes nos dois idiomas. *Save/leave room for dessert* pode ser considerada equivalente para 'Reserve espaço para a sobremesa'. O produto final será um glossário de livre acesso que auxilie tradutores e redatores da área a redigir textos que fluam com naturalidade na língua de chegada, por meio de uma abordagem empírica baseada em textos autênticos, considerando-se diferenças linguísticas e culturais nas sociedades envolvidas.

93

PROPOSTA DE UM VOCABULÁRIO BILÍNGUE DE FESTAS POPULARES BRASILEIRAS BASEADO EM UM ESTUDO DE *CORPUS*

Giovana Martins de Castro Marqueze
Programa de Pós-Graduação em Estudos da Tradução – Universidade de São Paulo (USP)
giovana.marchese@hotmail.com

Introdução: Este resumo trata de uma pesquisa em andamento que visa propor uma obra terminográfica bilíngue na direção português/inglês sobre festas populares brasileiras que tenha como público-alvo profissionais que produzem textos turísticos em inglês sobre o Brasil, sejam eles traduções ou produções nesse idioma. A proposta da obra é orientada pela análise de um *corpus* paralelo e de um *corpus* comparável bilíngues contendo textos turísticos. O intuito desse estudo é observar os desafios encontrados pelos tradutores e redatores desse gênero textual ao depararem com os referentes culturais relativos a festas populares brasileiras para que, assim, possamos pensar uma obra que vá ao encontro de suas necessidades.

Metodologia: Para esta pesquisa, que tem a Linguística de Corpus como aporte metodológico, foram compilados um *corpus* paralelo e um *corpus* comparável bilíngues. O primeiro é composto por textos turísticos

sobre o Brasil escritos originalmente em português com tradução para o inglês e o segundo é composto por textos turísticos sobre o Brasil escritos originalmente em português e por textos escritos originalmente em inglês. Os textos em português do *corpus* comparável foram submetidos à análise semiautomática com a ferramenta Keywords do programa AntConc 3.4.4 para extração dos termos relativos a eventos populares brasileiros, os quais serão analisados e formarão as entradas da obra terminográfica proposta. Trinta e cinco termos foram selecionados e divididos em cinco categorias: Festas populares: Carnaval, Bumba-meu-boi/Boi-bumbá, Festival Folclórico de Parintins, Festa Junina/ Festa de São João, Festa do Divino, Círio de Nazaré, Procissão do Fogaréu, Cavalhadas. Músicas e Danças: samba, samba-enredo, forró, xote, xaxado, frevo, marchinha, axé, tambor-de-crioula, quadrilha. Símbolos: (Boi) Caprichoso, (Boi) Garantido, boneco gigante de Olinda, arraial. Organização: escola de samba, trio elétrico, bloco de rua/de carnaval, bumbódromo, sambódromo, abadá. Comidas típicas: tapioca, pamonha, canjica, quentão, pé de moleque. Esses termos estão tendo suas traduções cotejadas no AntPconc 1.1.0, concordanciador de *corpora* paralelos. Concomitantemente, está sendo explorado o *corpus* comparável a fim de auxiliar na elaboração de definições dos termos na obra terminográfica proposta.

Resultados parciais: Os resultados obtidos até o momento a partir da análise do *corpus* paralelo mostram que o mesmo texto turístico escrito originalmente em português traz diferentes traduções e definições para um mesmo referente cultural, o que muitas vezes deixou o texto em inglês confuso, podendo prejudicar um dos objetivos desse gênero textual, que é o de atrair turistas para o local e/ou atividade promovidos. A análise do *subcorpus* comparável em inglês do termo "Festa de São João" mostrou que não está claro para os autores de dois dos três guias estrangeiros que a festa faz parte das nossas chamadas "Festas Juninas". Enquanto um autor afirma ser o "São João" uma festa de Candomblé, o outro informa que a festa trata de um concurso de dança no Nordeste. Esse resultado reforça a importância de uma obra terminográfica na área que contenha informações enciclopédicas. O modelo de verbete está em fase de elaboração.

96

EXTRAÇÃO E ANÁLISE DE PADRÕES LÉXICO-GRAMATICAIS EM ARTIGOS ACADÊMICO-CIENTÍFICOS EM DIFERENTES ÁREAS

Cristiane Kilian e Sandra Loguercio
Universidade Federal do Rio Grande do Sul (UFRGS)
ckkilian@cpovo.net, sandraloguercio@hotmail.com

Para a redação de artigos acadêmico-científicos é necessário, além de empregar adequadamente o léxico e as fraseologias de conhecimento especializado da área, também fazer uso das convenções retóricas utilizadas nesse gênero discursivo. Como contribuição para o estudo desses modos de dizer, o objetivo deste trabalho é apresentar a extração e a análise de padrões léxico-gramaticais da linguagem científica empregada em artigos acadêmico-científicos relacionados ao gênero e não propriamente à área de especialidade. Como *corpus* de estudo, utilizamos artigos em língua portuguesa de três áreas distintas: Linguística, Engenharia de Materiais e Ciências Econômicas, cole-

tados de periódicos indexados no SciELO. Procedemos à extração e análise segundo os aportes teórico-metodológicos da Linguística de Corpus. Utilizamos as ferramentas de análise linguística do AntConc - versão 3.4.4 (Anthony, 2014) para a geração de listas de palavras, geração de n-gramas e *clusters*, bem como geração de concordâncias. Com base em Tutin (2007), classificamos as estruturas encontradas em metacientíficas, aquelas relativas ao fazer científico (A análise dos dados mostrou que...; Este trabalho tem como objetivo...) e metadiscursivas, aquelas relativas à interação autor e destinatários, às relações internas do texto, ao posicionamento do autor, etc. (Nas seções a seguir,...; Vale ressaltar o papel fundamental de...). Quanto à macroestrutura, a análise aponta para diferenças na organização textual dos artigos das três áreas de especialidade analisadas, sendo os artigos da área de Engenharia de Materiais os que apresentam uma estrutura mais fixa e com menos variação. Quanto aos padrões léxico-gramaticais, verificamos que as estruturas encontradas podem estar vinculadas a uma parte específica do texto ou ocorrer em mais de um segmento textual, não estando relacionadas à sua macroestrutura, mas à argumentação e à construção discursiva de cada parte, que exige muitas vezes os mesmos movimentos retóricos (cf. Swales, 1990). Os resultados deste trabalho darão subsídios para a construção de objetos de aprendizagem (OA) online e de livre acesso que contribuirão para o desenvolvimento de competências e habilidades relativas à leitura, compreensão e produção de textos no meio acadêmico.

98

TRADUÇÃO AUTOMÁTICA: VILÃ OU NEM TANTO?

Michel de Sousa Teixeira e Cláudia Freitas
Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO)
mteixeira@aluno.puc-rio.br, claudiafreitas@puc-rio.br

A tradução automática (TA) é, hoje, uma ferramenta amplamente utilizada não só por empresas de serviços de tradução e tradutores profissionais, mas também por pessoas comuns, para diversos objetivos. Este trabalho analisa, em termos quantitativos e qualitativos, o desempenho de ferramentas de TA e de tradução assistida por computador (CAT) disponíveis comercialmente. Busca, também, apontar as perspectivas de uso da TA no futuro.

101

O USO DA LINGUÍSTICA DE CORPUS E *CORPORA* DE APRENDIZES PARA A PRODUÇÃO DE ATIVIDADES DIDÁTICAS E PARA O ENSINO DE INGLÊS COM FINS ACADÊMICOS

Anna Luisa Lopes Alves
Departamento de Letras Modernas –
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)
annalopes.alves@gmail.com

No mundo globalizado em que vivemos, notamos que a utilização da língua inglesa (doravante LI), tida como língua franca (Jenkins, 2007), é prática recorrente em

meio acadêmico, devido ao progresso da internacionalização de universidades do Brasil, assim como a demanda de publicações em revistas científicas por alunos de instituições acadêmicas brasileiras. Utilizaremos, portanto, a área de estudos da Linguística de Corpus para embasar a presente pesquisa quanto à utilização de ferramentas linguísticas e tecnológicas e o uso de *corpus* para o ensino de LI (Berber Sardinha, 2010; Viana, 2010) e também para o desenvolvimento de atividades didáticas para um curso de Inglês com Fins Acadêmicos. De acordo com a produtividade do uso de *corpora* para o ensino e para tradução pedagógica, podemos dizer que a Linguística de Corpus (LC) coopera para o ensino que se desenvolve a partir da descrição da linguagem, e essa descrição que se apoia nos *corpora* de aprendizes teve um impacto considerável sobre o ensino de línguas estrangeiras, portanto, as informações anteriores justificam nossa escolha pelo uso da LC e de *corpus* de textos escritos para o desenvolvimento de atividades pedagógicas (ainda em fase de desenvolvimento) que visam o ensino e aprendizagem mais autônomos dos alunos. Os objetivos dessa pesquisa são: mostrar como a LC, utilizada em um curso de Inglês com Fins Acadêmicos (Swales e Feak, 2009; Dayrell, 2011; Charles, 2012; Pinto, no prelo), é efetiva quando se trata do desenvolvimento da escrita de resumos de artigos (*abstracts*); analisar estruturas e palavras relevantes no *corpus* de aprendizes que caracterizam a escrita acadêmica em inglês e, desenvolver atividades didáticas pedagógicas, a partir da análise dos textos, para futuros cursos de língua inglesa e Inglês com Fins Acadêmicos baseados em *corpora*. Essa pesquisa está em fase de desenvolvimento e, para alcançarmos nossos objetivos utilizaremos resumos de artigos escritos em língua inglesa, por alunos não nativos de graduação e mestrado da Unesp de São José do Rio Preto, para a compilação do *corpus* de aprendizes. Após recolher os dados dos alunos, usaremos a ferramenta computacional de pesquisa linguística, o AntConc®, para desenvolvermos uma análise que busca estabelecer a comparação da escrita de resumos de artigo em inglês por alunos nativos (cujos textos fazem parte do *corpus* do Michigan Corpus of Upper-Level Sudents Papers – MICUSP) e não nativos (alunos da Unesp). A comparação levará em conta as palavras que exprimem "urgência" e "importância" (palavras, estas, que cativam a atenção do leitor para que se interesse pelo trabalho apresentado pelo aluno), como estas são usadas, o contexto, o cotexto e as estruturas gramaticais em que as palavras pesquisadas estão inseridas nos textos acadêmicos dos aprendizes. Ademais, compararemos como algumas palavras essenciais para um texto acadêmico (como *text markers* e palavras gramaticais, por exemplo) são utilizadas na prática de escrita acadêmica dos alunos nativos e não nativos. Como resultados, criaremos listas de concordâncias das palavras a serem analisadas e com essas listas desenvolveremos atividades didáticas para serem colocadas em prática em cursos de Inglês com Fins Acadêmicos para as três grandes áreas do conhecimento, e assim, contribuir para o desenvolvimento da aprendizagem de LI na comunidade acadêmica brasileira.

TEXTUAL DISCOURSE ANALYSIS: TOWARDS A PRAGMATIC-ARGUMENTATIVE MODEL FOR THE INTERNATIONAL LEGAL DISCOURSE

Hális Alves do Nascimento França
Programa de Pós-graduação em Estudos da Linguagem (PPgEL) – Universidade Federal do Rio Grande do Norte (UFRN)
halisalves.uni@gmail.com

Introduction: This research proposal concerns the development of a theoretical model of pragmatic-argumentative nature, having as starting point the Textual Discourse Analysis and, as object, the international legal discourse in English language, i. e., multilateral environmental agreements (MEAs). For that purpose, it is fundamentally grounded on branches of theoretical and descriptive linguistics, focusing further on pragmatics and argumentation theory. The purpose of the model lies in the corpus-based identification of the acting forces underlying selected sentences from MEAs, considering the measure or degree to which that given sentence means to enforce a given legal commitment. For example, when a certain sentence states that (a) Parties shall provide, whereas another states that (b) States shall take all practicable steps to promote, what kind of forces act on those statements, in such a way as to define their measurement or degree of enforcement? Is (a) stronger than (b), in the sense that the Parties become more committed? If so, how does it happen and how frequently so? In principle, this research suggests that these forces are structured around an interaction that considers the degree of strength of illocutionary forces (grounded on speech act theory and illocutionary logic analytic tools) and the force of their arguments (grounded on argumentation theory, legal argumentation and informal logic). Methodology: The model follows the Textual Discourse Analysis on its final level of examination (speech acts and argumentative orientation) and consists of several procedures comprised in five stages: (1) determination of global features of the corpus and text plan; (2) selection of sentences compatible to the scope of the model, i. e., those that reflect any measure or degree of enforcement; (3) bi-operational analysis of illocutionary and argumentative forces acting on those sentences; (4) analysis of the correlation between the resulting illocutionary-argumentative forces and subsequent mapping of their occurrences and interactions throughout the corpus. To the extent of the obtained data therewith, the model will proceed to the (5) development of an intensity scale able to pinpoint the measure or degree of enforcement of those forces. According to their distribution in an intensity continuum, ranging from 1 (evidently/absolutely binding) to < 1 (increasingly less binding), the model will position corpus data in such a way that it may provide a different display for each MEA, indicating, thus, how strong are the obligations contained in their texts, along with detailed information on each commitment by means of their pragmatic/argument maps. Partial Results: Preliminary results are still forthcoming. However, it is generally expected that this model will contribute to the description of linguistic phenomena related to the international legal discourse and aid in a number of problems concerning legal hermeneutics and formalization of legal arguments. Moreover, the author intends to foster communi-

cation with researchers on corpus linguistics and computational linguistics in order to improve the analytic model currently being developed towards a more efficient tool of linguistic quali-quantitative description of natural languages in the context of international agreements.

104

LEGENDAGEM, *CORPORA* E ANÁLISE DO DISCURSO

Elaine Trindade
Universidade de São Paulo (USP)
elainetrindade@uol.com.br

O objetivo deste trabalho é desenvolver a Análise do Discurso de duas versões de legendas em português da série Suits, primeira temporada, visando identificar variações do discurso considerando a mensagem final dos textos em português, em comparação com o texto original em inglês. Com relação às duas versões, serão utilizadas a legenda oficial exibida pelo streaming de vídeo Netflix e as legendas feitas por fansubs que são disponibilizadas gratuitamente em sites na internet. As legendas de fansubs são consideradas por Diaz-Cintas e Munhoz (2006) como legendas feitas por amadores. Como Netflix contrata tradutores profissionais, a análise também apresenta uma comparação entre legendas profissionais e amadoras. A legendagem é uma área dos Estudos da Tradução que desperta um grande interesse na população em geral. Entretanto, quando buscamos estudos acadêmicos nesta área, é possível identificar que esses estudos ainda estão em suas fases iniciais, visto que eles foram iniciados há pouco mais de 20 anos. Há, principalmente em língua portuguesa, uma carência em estudos que abordem a qualidade da tradução quando consideramos a mensagem transmitida nas legendas se comparadas ao áudio original. Sendo assim, um estudo que visa desenvolver a análise do discurso de duas versões de legendas frente ao original, pode apresentar dados importantes para avanços na qualidade da tradução para legenda. Um ponto que vale destaque, é que o processo utilizado por profissionais e amadores são bem distintos. O profissional é obrigado a seguir regras e normas definidas pelos exibidores na composição do texto em língua portuguesa, enquanto o amador tem liberdade no processo tradutório. Ambos seguem as regras técnicas, mas no momento da tradução, as escolhas de palavras são mais livres para o amador. Para desenvolver a pesquisa, a Linguística de Corpus, utilizando o software AntConc, terá papel fundamental no levantamento de colocados e *keywords*. Após a identificação de dados conflitantes entre as duas versões, análises mais detalhadas com relação ao discurso são realizadas e as orientações para esse estudo seguem critérios apresentados por Fairclough (1995), Van Dijk (1980) e Baker (2006).

ELABORAÇÃO DE TAREFAS PEDAGÓGICAS COM BASE EM *CORPUS*: EXPLORANDO RECURSOS LINGUÍSTICOS

Anamaria Welp e Álvaro R. Didio
Departamento de Línguas Modernas – Instituto de Letras –
Universidade Federal do Rio Grande do Sul (UFRGS)
anamaria.welp@ufrgs.br, alvaro.didio@ufrgs.br

Este trabalho visa relatar o andamento do projeto de pesquisa intitulado Construção de um banco de tarefas pedagógicas para as disciplinas de inglês do curso de Letras: compilação de *corpus* e aplicações didáticas. O projeto tem por objetivos [1] compilar textos que circulam dentro das temáticas e gêneros do discurso abordados nas disciplinas de Inglês I a V do curso de Letras da Universidade Federal do Rio Grande do Sul (UFRGS) – cujo currículo é organizado por eixos temáticos – para então [2] elaborar tarefas pedagógicas que explorem os recursos linguísticos mais presentes em tais textos.

Na primeira etapa do projeto foi compilado um *corpus* para cada temática. A fase apresentada aqui teve a finalidade de [1] atualizar o *corpus* através de compilação de novos textos para cada disciplina; [2] analisar o *corpus* através do software AntConc na busca de padrões de uso da linguagem, ou seja, palavras-chave, palavras mais frequentes, agrupamento de palavras e colocações presentes nos gêneros e temáticas de cada disciplina; e, [3] como resultado, elaborar tarefas pedagógicas que visem a familiarização dos alunos com os recursos linguísticos necessários para a compreensão e produção de textos exigidas nas disciplinas.

A título de ilustração, será apresentada uma tarefa elaborada para uma sequência didática da disciplina de Inglês IV, cujo eixo temático é Contemporary Issues in Cinema and Literature e o produto final é uma resenha cinematográfica. O *corpus* para a elaboração desta tarefa, portanto, foi composto de 3.005 resenhas cinematográficas obtidas a partir do website IMDb (Internet Movie Database), resultando em 25.248 *word types* e 1.067.248 *word tokens*; também, quanto à extração dos padrões léxico-gramaticais mais estáveis e frequentes, foram selecionados aqueles de 4 e 5 palavras e com frequência mínima de 12 ocorrências. Estes dados foram a base para a criação da tarefa em questão e permitem que os alunos produzam seus textos a partir do uso real da língua evidenciado pela análise do *corpus*.

COMENTCORPUS: O USO DE MECANISMOS LINGUÍSTICOS NA DETECÇÃO DE IRONIA E SARCASMO PARA O PORTUGUÊS DO BRASIL EM UM *CORPUS* OPINATIVO

Gabriela Wick Pedro e Oto Araújo Vale
Universidade Federal de São Carlos (UFSCar)
gabiwick@gmail.com, otovale@gmail.com

A ironia e o sarcasmo são mecanismos que são capazes de alterar o sentido de uma sentença ou transformar sua polaridade positiva para negativa e vice-versa. O estudo das expressões irônicas pode ser um bom recurso no sentido de aprimorar ferramentas de Análise de Sentimentos, ou Mineração de Opinião, domínio do Processamento de Linguagem Natural, que tem como objetivo analisar computacionalmente opiniões, emoções, sentimentos e subjetividades presentes em textos (Liu, 2012). Pesquisas para o inglês (Gonzáles-Ibáñez et al., 2011), para o português europeu (Carvalho et al., 2009) identificaram padrões e mecanismos linguísticos que podem detectar sentenças irônicas em *corpora*. Freitas (2013) apresentou alguns desafios que a ironia apresenta para a identificação de opinião em textos do português do Brasil. No presente projeto busca-se analisar e descrever mecanismos subjacentes à expressão de ironia em textos avaliativos no português do Brasil, tendo como propósito a criação de formalizações capazes de detectar possíveis construções irônicas e sarcásticas. Para examinar as ocorrências de ironia e sarcasmo partiremos da busca em um *corpus* previamente construído por comentários de notícias e anotado de acordo com as expressões irônicas e sarcásticas identificadas. O *ComentCorpus* é um *corpus* de comentários de notícias anotado manualmente com informações de sentimento. A coleção é composta por 6.185 comentários do domínio político retirados de 90 notícias relacionadas ao período pré-impeachment do caderno Poder do jornal Folha de São Paulo, totalizando 14 mil sentenças e 207 mil palavras. A anotação do *corpus* é baseada na definição de opinião proposta por Liu (2012), onde toda opinião é composta por pelo menos dois elementos fundamentais: alvo (podendo ser uma entidade, aspecto de uma entidade, um produto, pessoa, organização, marca, evento, etc.) e um sentimento (que representa uma atitude, opinião ou emoção em que o autor da opinião tem a respeito de um determinado alvo). Até o presente momento o *corpus* foi anotado de acordo com o autor da opinião, data, alvo da opinião e presença eventual de ironia. Cada comentário contém um autor de opinião sobre um determinado alvo e a data na qual a opinião foi expressa pelo autor. Uma sentença opinativa pode conter expressões irônicas subdivididas previamente em diminutivo, aspas e pontuação. Após a primeira etapa com a identificação de expressões, nas etapas seguintes inserem-se a anotação sentenças irônicas de acordo com mecanismos utilizados para expressar ironia e a classificação em diferentes categorias, representados em padrões léxico-sintáticos. Assim, espera-se estabelecer uma lista com classificações e características de ironia e sarcasmo, colaborando com o desenvolvimento da área e aperfeiçoando ferramentas de identificação automática de opinião.

LEXICAL BUNDLES ACROSS THE METHODOLOGY AND RESULTS SECTIONS OF APPLIED LINGUISTICS RESEARCH ARTICLES

Luciana de Macedo
Universidade Federal de Minas Gerais (UFMG)
lucianadiasdemacedo@gmail.com

Bearing in mind the relevance of appropriate use of lexical bundles in the academic discourse, this research provides a closer analysis of how researchers employ these elements, in connection with the rhetorical moves scheme, in specific sections of their research articles (RA) in Applied Linguistics. Lexical bundles and rhetorical moves are both considered building blocks to be used in the construction of discourse (Cortes, 2013). Lexical bundles have been defined as recurrent expressions that can be retrieved from our memory as "text building blocks" (Biber, Conrad e Leech, 2002, p. 443) and "important building blocks of coherent discourse and characteristic features of language use in particular settings" (Hyland, 2008, p. 8). Similarly, rhetorical moves are explained as "main building blocks" of a genre (Biber, Connor and Upton, 2007, p. 53). Dudley-Evans (1995) also highlights that rhetorical moves are so inherently part of a genre that they should be used to teach novice writers how to produce successful texts in a particular genre. For that reason, we have begun to compile a list of the most frequently used lexical bundles, connected to the rhetorical moves scheme from two sections of Applied Linguistics RAs: 1. Methodology and 2.Results/Discussion. The RAs analyzed were written by native English speakers and/or highly proficient English speakers. In order to retrieve reliable and valid data, each subcorpus *presents a significant number of words, roughly 300,000. They were all collected via Periódicos Capes, from six major areas of Applied Linguistics, so as to increase the* corpus *representativeness. The population of the* corpora *was selected based on the definition of areas incorporated in the 2016 AILA Congress (International Association of Applied Linguistics). After this process of compilation, the lexical bundles were generated and analyzed in consonance with a move scheme designed by Swales (1990, 2004), and its adaptations (Yang e Allison, 2003; Cortes, 2004; Kanoksilapatham, 2005 and Cortes, 2013). Primary results show that the Discussion/Results section yields a high number of lexical bundles which communicate stance, such as appear(s) to be; at least in; is likely to (be); to some extent; along many others. This corroborates what Hyland (1998) claimed about the process of persuasion. He states that persuading depends on how well experiments and results are presented and contextualized, this should be done with caution, precision and humility. The lexical bundles identified in the Discussion/Results section imply a close relationship with the moves proposed by Kanoksilapatham's (2005), such as in Results: Providing justifications for the choice of procedures, Stating results, and Commenting results; and in Discussion: Consolidating results, Stating limitations, and Suggesting further studies. The next steps are to understand what other lexical bundles are typical of each section separately; investigate whether there is a relationship between the bundle structure and its function. Finally, our ultimate aim is to create a comprehensive list of the bundles within each category of moves or steps from each section analyzed.*