CADERNO DE RESUMOS

COMUNICAÇÕES ORAIS

ANÁLISE DA FREQUÊNCIA DE OCORRÊNCIA EM PEQUENOS CORPORA: UMA PROPOSTA METODOLÓGICA DE COMPARAÇÃO DE CORPORA DISTINTOS

Alan Jardel de Oliveira (UFMG)

Este trabalho apresenta uma proposta metodológica de comparação de corpora distintos, com a finalidade de averiguar se há correlação entre a freqüência das palavras em corpora de fala do português brasileiro coletados em regiões distintas e entre corpora orais e escritos coletados em uma mesma região. O estudo faz parte do grupo de pesquisa VARFON-Minas (CNPQ), que tem como objetivo, entre outros, descrever e analisar os aspectos fonéticos, fonológicos, morfológicos, lexicais, sociolinguísticos e da formação sócio-histórica dos dialetos mineiros. Em Oliveira (2006) apresentamos uma análise da variação da sílaba final átona composta por consoante lateral alveolar mais vogal no falar de Itaúna. Neste estudo foram identificadas quatro formas variantes: a sílaba plena, o apagamento da vogal, a velarização da lateral após apagamento da vogal e o apagamento da sílaba. O objetivo da pesquisa foi identificar e analisar os fatores que favorecem a realização de tais variantes lingüísticas. O trabalho teve como base teórica e metodológica a sociolingüística variacionista, desenvolvida por Willian Labov a partir da década de 60. Nessa perspectiva, considera-se que a língua é um sistema heterogêneo e a variação e a mudança linguística são inerentes a esse sistema (cf. Labov, 1972). Assim, a produção de uma variante lingüística está associada probabilisticamente a algum fator linguístico ou social e, portanto, não é aleatória. Um dos fatores que podem ser levados em consideração na análise da variação e da mundança lingüística, conforme Bybee (2001) é a freqüência de ocorrência dos itens lexicais. É possível que certos processos variáveis sejam influenciados pela freqüência de ocorrência de uma palavra. Assim, seria importante averiguar se palavras mais freqüentes realizam-se mais de uma forma variante do que de outra, o que seria indício de que a freqüência da palavra interfere na forma como ela é produzida foneticamente. Porém, como determinar se freqüência observada em uma amostra corresponde à freqüência na língua? Sardinha (2000) afirma que a amostra de corpora de linguagem deve ser a maior possível de forma que ela seja mais aproximada da população do qual ela deriva. A metodologia de coleta de dados proposta pela sociolingüística variacionista (cf. Labov (1972)) prevê que a análise seja feita por meio da coleta de dados reais de fala, coletados na comunidade de fala no qual o processo variável em análise está ocorrendo. Pelas limitações impostas pela pesquisa realizada em Itaúna (tempo, recursos financeiros, etc.), a quantidade de palavras coletadas nos dados de fala espontânea não constitui-se em um corpus de tamanho significativo. Nesse corpus, foram coletadas pouco mais de 76 mil palavras, o que, de acordo com Sardinha (2000), constituiria-se como um corpus pequeno (para o autor, com até 80

mil palavras o corpus seria classificado como pequeno; corpora médio-grandes seriam aqueles que apresentassem mais do que 1 milhão de palavras). Assim, devido à limitação do tamanho da amostra de dados de fala em Itaúna, a classificação dos itens como muito freqüentes ou pouco freqüentes perde confiabilidade. Por outro lado, existem diversos corpora do português brasileiro falado de tamanhos mais significativos já coletados e disponibilizados para consulta via internet. Entretanto, não sabemos se a freqüência das palavras entre corpora de regiões distintas podem ser de fato comparados. Da mesma forma, não sabemos se corpora de fala e corpora de escrita podem ser comparados. Nesse estudo, apresentamos uma metodologia de comparação de corpora distintos. Para tal, foram analisados e comparados três corpora: (1) um corpus de fala espontânea coletado na cidade de Itaúna/MG composto de 76.027 palavras; (2) um corpus de escrita, coletado dos três principais jornais da cidade de Itaúna/MG composto de 2 milhões de palavras e (3) um corpus de fala do LAEL (PUC/SP), composto de aproximadamente 3 milhões de palavras. Para testar a possibilidade de comparação entre as freqüências dos itens observadas nos diferentes corpora foi criado um banco de dados no qual foram inseridas todas as palavras analisados em Itaúna (terminadas em sílaba átona composta por lateral alveolar mais vogal). A cada uma dessas palavras foi associada a quantidade de vezes que eles apareciam em cada um dos três corpora utilizados, criando-se assim 3 variáveis contínuas. A partir desse banco de dados, analisamos a correlação entre as variáveis utilizando um estimador chamado coeficiente de correlação de Spearman. Tal coeficiente cria uma espécie de ranqueamento dos dados, minimizando o efeito dos valores atípicos, bastante comuns nesse tipo de estudo. Após a análise, concluímos que há uma correlação estatisticamente significativa entre os corpora de fala de Itaúna, de fala do LAEL e de escrita de Itaúna/MG e que, portanto, a freqüência das palavras observada nos corpora de tamanhos mais significativos pode servir de referência para se estabelecer a freqüência das palavras observadas no corpus de fala da cidade de Itaúna/MG, de tamanho menos significativo.

Contato: alanjardel@gmail.com

DESAFIOS PARA A ANOTAÇÃO SEMÂNTICA DE TEXTOS JURÍDICOS: LIMITES NO USO DA FRAMENET E ROTAS ALTERNATIVAS

Anderson Bertoldi (Unisinos)
Rove Luiza De Oliveira Chishman (Unisinos)

Diferentes projetos têm utilizado a FrameNet para anotação de textos em diferentes línguas. Apesar de as etiquetas semânticas da FrameNet serem utilizadas para a anotação de textos em diferentes línguas, essas etiquetas foram criadas a partir da análise de unidades lexicais em inglês. Este trabalho mostra que o uso das etiquetas semânticas da FrameNet pode ocasionar algumas divergências no que tange a sua aplicação a outras línguas para anotação de textos de domínios socialmente orientados, como o Direito. A FrameNet (Fillmore et al., 2003) descreve o significado das palavras através da conexão de cada palavra a um frame semântico. Segundo Fillmore et al. (2003), as unidades primárias de descrição lexical na FrameNet são o frame e a unidade lexical. O frame, segundo Fillmore (1982), é um sistema de conceitos relacionados de tal forma que, para se entender um conceito, é necessário a compreensão de todos os conceitos relacionados. A unidade

lexical é compreendida como a soma de uma palavra e um significado. Na FrameNet, a informação sobre valência é especificada em dois níveis: o sintático e o semântico. A valência sintática especifica os tipos sintagmáticos (sintagma nominal, preposicional etc) e as funções gramaticais (sujeito, objeto etc). A valência semântica é descrita em termos de entidades que podem participar de um frame evocado por uma unidade lexical, tais entidades são chamadas de "elementos de frame" (Fillmore et al., 2003). O projeto SALSA (Saarbrücken Lexical Semantics Annotation and Analysis) vem realizando a anotação de corpora em língua alemã a partir das etiquetas semânticas da FrameNet (Burchardt et al., 2009). O projeto SALSA parte da suposição de que é possível reutilizar as etiquetas semânticas da FrameNet para a análise semântica do alemão. Assim, esse projeto inclui (i) a anotação de um corpus de grande porte em alemão e a geração de léxico baseado em frames a partir da anotação do corpus e (ii) a indução de modelos baseados em dados para análise semântica automática e aplicações em processamento de linguagem natural (Burchardt et al., 2009). O trabalho do projeto SALSA com a anotação de corpora de grande extensão, geração automática de entradas lexicais a partir de corpora anotados e análise semântica automática influenciou vários trabalhos de criação automática de FrameNet. A proposta de Padó e Lapata (2005) sugere o uso de corpora paralelos para a criação automática de entradas lexicais baseadas em frames. A partir da anotação de um corpus em inglês com as etiquetas da FrameNet, seria possível transferir a anotação do corpus em inglês para um corpus de outra língua. Essa técnica vai inspirar vários trabalhos de transferência automática de anotação semântica, como Padó e Pitel (2007), para o francês, Tonelli e Pianta (2008), Dini e Bosca (2009) e Venturi et al. (2009), para o italiano. Este trabalho objetiva analisar a aplicabilidade das etiquetas semânticas da FrameNet para anotação de corpora, considerando-se a variante brasileira e o sistema judiciário brasileiro. Para alcançar esse objetivo, as unidades lexicais evocadoras do frame Criminal_process foram contrastadas com seus equivalentes em português. São estudadas as unidades lexicais evocadoras de onze subframes que compõem o frame Criminal_process. Foram analisados (i) a equivalência das unidades lexicais entre o inglês e o português e (ii) o contexto jurídico (o frame) evocado por cada unidade lexical em inglês e em português. O estudo contrastivo demonstrou que os frames semânticos em domínios socialmente construídos, como o Direito, apresentam um alto grau de divergência entre as línguas. A diferença entre os sistemas jurídicos norte-americano e brasileiro faz com que os eventos jurídicos não sejam os mesmos nos Estados unidos e no Brasil. Isso provoca uma quebra entre a equivalência das unidades lexicais e a correspondência dos eventos jurídicos. Algumas unidades lexicais do inglês que apresentam equivalência em português podem não apresentar correspondência conceitual, ou seja, o significado é semelhante, mas o evento jurídico descrito por essa unidade lexical em inglês não é totalmente semelhante em português. O estudo contrastivo apontou para a necessidade de se criarem frames jurídicos específicos para o sistema jurídico brasileiro. O tratamento da equivalência de frames é relativamente recente (Lönneker-Rodman, 2007). Em geral, os estudos de equivalência se detêm no estudo da equivalência de unidades lexicais. Os frames jurídicos são evidências da falta de equivalência de frames entre línguas, uma vez que esses frames representam um conhecimento socialmente construído, sendo, portanto, específicos de cada cultura e de cada país. A partir do estudo contrastivo dos frames semânticos da FrameNet e dos frames criados para o processo penal brasileiro, percebe-se que os frames jurídicos apresentam diferentes níveis de equivalência, variando desde a quantidade de elementos de frame até a natureza do evento jurídico descrito pelo frame. Enquanto o par de frames Try_defendant e Julgar_acusado possui unidades lexicais equivalentes, descrevem um evento jurídico semelhante e

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

Contato: andersonbertoldi@yahoo.com

CORPOS E CORES: COLORINDO A DESCRIÇÃO DA LÍNGUA PORTUGUESA

Claudia Freitas (PUC-Rio)
Diana Santos (Universidade de Oslo)

O estudo das cores ocupa um papel importante no debate sobre universalismo e relativismo, interessando a diferentes áreas do conhecimento. Como possui aspectos biológicos e linguísticos, é natural que o campo das cores seja de especial relevância nos estudos da/sobre a linguagem. No entanto, como notam Santos et al. (2011), boa parte dos estudos sobre a cor, defendendo a universalidade da conceptualização da cor (Berlin & Kay, 1969; Rosh, 1975) ou, pelo contrário, refutando-a (Wierzbicka, 1990), tomam como base experiências com informantes. Como também aponta Lucy (1997), as pesquisas sobre as cores têm se concentrado na comparação entre as línguas, em um refinamento da tipologia e no reforço de argumentos de base biológica, enquanto relativamente pouco tem sido feito para melhorar a qualidade da descrição linguística. Nesse contexto, trabalhos com base em corpos podem oferecer um bom complemento no que se refere ao comportamento das cores nas línguas. Com relação ao português, o trabalho de Biderman et al. (2007), assim como o nosso, também parte de corpos para explorar e descrever as diferenças na expressão da cor no Brasil e em Portugal. No entanto, diferentemente do trabalho aqui apresentado, o estudo toma por base apenas o lema de algumas cores (azul, vermelho e encarnado), não considerando a possibilidade de se considerarem grupos de cor, bem como os vários sentidos que as palavras de cor podem assumir, não necessariamente diretamente vinculados a cor (como “vinho verde”; “sangue azul”; “imprensa marrom”; “chapa branca”). Neste trabalho, damos continuidade à exploração das cores em corpos da língua portuguesa, ao mesmo tempo em que, aproveitando as possibilidades de pesquisa em corpos das variantes do Brasil e de Portugal, contrastamos também o uso nas duas variantes. De maneira geral, este trabalho pode ser entendido como uma continuação de XXX, que, ao apresentar dados que sugerem um maior uso das cores na variante portuguesa, levanta questões sobre o uso das cores no Brasil e em Portugal. Para tanto, partindo de uma investigação sobre a distribuição das categorias gramaticais nas duas variantes, comparamos, em um primeiro momento, a distribuição das cores por categoria gramatical. Além disso, tirando proveito de corpos ricamente anotados com informação linguística, exploramos a distribuição das cores por categoria sintática (e por variante). Por fim, ao aprofundarmos os achados de XXX quanto aos grupos de cores mais constantes, por um lado, e mais variáveis, por outro, entre Brasil e Portugal, buscamos contribuir para um quadro descritivo

mais completo das cores na língua portuguesa, bem como das diferenças entre as variantes. O trabalho de exploração de corpos toma por base o projeto AC/DC (Acesso a Corpos / Disponibilização de Corpos), que disponibiliza corpos do português na internet (Costa et al, 2009). O AC/DC contém atualmente cerca de 375 milhões de palavras, distribuídas em cerca de 16 milhões de frases, de diferentes gêneros e nas variantes do Brasil e de Portugal. Todo o material foi anotado automaticamente pelo PALAVRAS (Bick, 2000), e algumas partes passaram por revisão humana. Além da anotação do PALAVRAS, os corpos do AC/DC também vêm recebendo anotação relativa à informação semântica no campo das cores (Mota & Santos, 2009; Frankenberg-Garcia & Santos, 2002; Inácio et al., 2008). No presente trabalho, ainda que a exploração das cores utilize dados de diversos corpos do AC/DC, a análise toma por base principalmente os corpos CONDIVport e CHAVE. O CONDIVport, criado com o objetivo de permitir o estudo da convergência e divergência do português entre as variantes do Brasil e de Portugal, contém cerca de 5 milhões de palavras distribuídas em jornais desportivos e revistas de saúde e de moda – áreas que tendem a empregar cores de maneira recorrente. Além disso, é importante mencionar que toda a anotação das cores no CONDIV passou por revisão humana. O CHAVE contém cerca de 99 milhões de palavras distribuídas em textos jornalísticos da Folha de São Paulo (Brasil) e do jornal Público (Portugal). Por ser um corpo maior, e de conteúdo mais geral, oferece dados complementares aos obtidos no CONDIVport. De maneira geral e resumida, a partir das questões levantadas no trabalho, podemos afirmar que: (i) a distribuição das categorias gramaticais é a mesma entre as variantes portuguesa e brasileira; (ii) a distribuição dos verbos de cor (por exemplo, amarelar; embranquecer) e dos adjetivos de cor também é a mesma entre as variantes portuguesa e brasileira; (iii) como esperado, verbos se prestam pouco a representar o processo de "colorização" das coisas. Assim, não é surpresa que a grande maioria dos verbos de cores esteja no particípio. E, quando consideramos apenas as formas finitas dos verbos, independente de variante ou gênero, certos grupos de cores, como castanho, creme, prateado e azul, desaparecem. Por outro lado, outros grupos, como verde, branco, preto, amarelo e dourado, são mais produtivos na formação de verbos, dando origem a dois ou mais lemas distintos; (iv) de uma perspectiva do uso, é interessante perceber o que fazem os verbos de cor. Branquear e embranquecer, por exemplo, ambos do grupo branco, têm comportamentos diferentes não apenas quanto à frequência. Branquear, mais frequente, é usado principalmente de maneira metafórica: muito mais que os dentes, branqueamos dinheiro e imagem. Embranquecer tem um uso mais literal, - embranquecemos os cabelos, a pele; (v) com relação aos adjetivos de cores, notamos que os grupos em que há a maior diferença entre as variantes são também aqueles em que o vínculo com a propriedade de colorir está mais distante: "conta laranja"; "deputado laranja"; "eminência parda"; "molho pardo", por exemplo. Por fim, reforçamos a ideia de que para comparar termos de cor entre duas ou mais línguas é importante uma caracterização detalhada das cores em cada uma das línguas contrastadas. Em nosso estudo, buscamos conjugar as dimensões de uso e de forma, oferecendo um quadro interessante das cores em português, considerando também distinções entre as variantes do Brasil e de Portugal.

Contato: maclaudia.freitas@gmail.com

ENSINO DE LÍNGUA ESTRANGEIRA COM CORPORA PARA PROFISSIONAIS DE PUBLICIDADE: UMA ABORDAGEM PARA O PROFESSOR

Cristina Acunzo (PUC-SP)

Este trabalho teve como objetivo o desenvolvimento de aulas de língua inglesa como Língua Estrangeira com o uso de corpora para profissionais da área de Publicidade. O ensino Línguas Estrangeiras para profissionais de áreas específicas é um desafio, pois existe pouco material disponível no mercado que atenda às necessidades dos alunos de comunicar-se em seu meio profissional. Além disso, o uso de corpora na preparação de materiais e sua aplicação na sala de aula ainda são incipientes; são poucos os materiais de ensino feitos com base em corpora e professores encontram dificuldades em criar seus próprios materiais utilizando um corpus, ferramentas e programas como, por exemplo, concordanciadores (BERBER SARDINHA, 2011). Para atingir nosso objetivo, encontramos suporte teórico principal na Linguística de Corpus, que proporciona a pesquisa, o estudo e a exploração da língua em uso (BERBER SARDINHA, 2004). Mais especificamente, o trabalho fundamenta-se na área de pesquisa baseada em corpus que se preocupa com o ensino de Línguas Estrangeiras. Considerando o contexto específico de aulas em agências de Publicidade e as dificuldades enfrentadas por professores de línguas estrangeiras, como falta de tempo para preparação de aula, altas cargas horárias e falta de material específico que auxilie o aluno a explorar e usar a linguagem de seu meio profissional, desenvolvemos uma abordagem para orientar a preparação das aulas com corpora. Essa abordagem baseia-se na proposta de Berber Sardinha (2011) com relação ao uso de corpora por meio de atividades centradas na concordância e centradas no texto e incorpora pressupostos como a criação de bancos de materiais com atividades criadas em um modelo (template), de forma que tenham partes intercambiáveis e sejam reutilizáveis. Com esse modelo, o professor pode preparar aulas que promovam a exploração de itens léxico-gramaticais da língua em uso por meio de materiais autênticos. No que concerne a autenticidade, baseamo-nos em Mishan (2005), que propõe uma abordagem para a exploração de textos autênticos no ensino de Língua Estrangeira, tanto em sala de aula, quanto na preparação de materiais didáticos. Mishan (2005) aborda duas questões relevantes para essa pesquisa: (1) critérios para a autenticidade na seleção de textos e na preparação de materiais para o ensino de Língua Estrangeira e (2) as bases pedagógicas e provenientes de pesquisa a favor do uso de materiais autênticos. Com essa pesquisa, buscamos preencher lacunas como a falta de pesquisa e aplicação de aulas com corpus, assim como no ensino de inglês para a área de Publicidade. Partimos das questões: (1) Quais os padrões lexicogramaticais distintivos do corpus de Publicidade? e (2) Quais atividades de ensino podem ser produzidas a partir desses padrões para o público-alvo? Para responder às questões, desenvolvemos a seguinte metodologia de pesquisa: (1) coleta de um corpus de 1 milhão de palavras composto por artigos escritos e transcrição de vídeos com textos lidos e entrevistas de uma revista digital de Publicidade; (2) análise do corpus e sua comparação com um corpus de referência, o BNC (British National Corpus) para a identificação dos padrões lexicogramaticais distintivos da área, por meio do programa Wordsmith Tools; (3) apresentação dos procedimentos de transposição dos achados para o ensino, ilustrando como preparar aulas e criar diversas atividades baseadas em corpora utilizando o modelo (template) proposto. Os resultados da análise nos permitiram propor atividades que proporcionam aos alunos a exploração de padrões como run ads, in social media e ad spending of por meio de textos e das ferramentas WordSmith Tools e

AntConc. Acreditamos que essa pesquisa contribui para o ensino de línguas estrangeiras e para a Linguística de Corpus de três formas: (1) no desenvolvimento de novos materiais para o ensino de Línguas Estrangeiras baseados em corpora, (2) na aplicação de abordagens que estão sendo desenvolvidas em sala de aula e (3) no ensino de língua inglesa para profissionais que atuam na área de Publicidade.

Contato: cristinaacunzo@hotmail.com

Referências:

BERBER SARDINHA, A. P. Linguística de Corpus. São Paulo: Manole, 2004.

BERBER SARDINHA, A. P. Como usar a Linguística de Corpus no ensino de língua estrangeira: por uma Linguística de Corpus educacional brasileira. In: VIANA, V. & TAGNIN, S. Corpora no ensino de línguas estrangeiras. São Paulo: Hub Editorial, 2011.

MISHAN, F. Designing Authenticity into Language Learning Materials. Bristol: Intellect Books, 2005.

CONTRIBUIÇÕES METODOLÓGICAS PARA O DESENVOLVIMENTO DA PLATAFORMA FRAMENET BRASIL: A DESCRIÇÃO DE ALGUMAS UNIDADES LEXICAIS DOS FRAMES FECHAMENTO E MOVIMENTO_CORPORAL

Gabriela da Silva Pires (Doutoranda / UFJF)
Maria Margarida Martins Salomão (UFJF)

Este trabalho vincula-se ao projeto de pesquisa de implantação do Projeto FrameNet Brasil e tem como objetivo empreender a descrição lexicográfica de sete Unidades Lexicais (ULs) que evocam a cena de abertura no frame Fechamento, a saber: quatro ULs monolexêmicas – desabotoar, desarrolhar, desatarraxar, destampar –; e três ULs polilexêmicas – abrir_((tampa)), levantar_((tampa)) e tirar_((tampa)). O respaldo teórico da pesquisa é a Semântica de Frames (FILLMORE, 1982; GAWRON, 2008; PETRUCK, 2008). Pretendemos que as análises feitas contribuam para a construção da contraparte para o português brasileiro da rede semântica FrameNet.

Contato: gabrielaniger@yahoo.com.br

AQUISIÇÃO DE LINGUAGEM ESCRITA, FONOLOGIA E ESTUDOS EM CORPORA: O PROJETO E-LABORE

Gustavo Mendonça (UFMG)
Thaïs Cristófaro-Silva (UFMG)
Leonardo Almeida (UFMG)

O presente trabalho tem por objetivo apresentar a fase atual em que se encontra o corpus do Projeto e-Labore. O e-Labore (Laboratório Eletrônico de Oralidade e Escrita) tem por propósito coletar, cadastrar e disponibilizar para a comunidade científica um banco de dados de material escrito por crianças de 6 a 12 anos. O objetivo central do trabalho é permitir um mapeamento do vocabulário infantil do português brasileiro atual, contribuindo, dessa forma, com os debates a respeito da interação entre a linguagem adulta e infantil, em um contexto de mudança linguística e evolução de linguagem. A Teoria de Exemplares (JOHNSON, 1997; PIERREHUMBERT, 2001) e a Fonologia de Uso (BYBEE, 2011) foram utilizadas como motivadores na elaboração do corpus. Atualmente, o projeto se encontra na fase de disponibilização dos dados: as fases de coleta e cadastro das redações já foram completadas. No total, duas coletas foram realizadas, o número de textos reunidos contabiliza 7817 (sendo 1952 textos referentes à primeira coleta, e 5865 à segunda). Todos os textos são de autoria de alunos de escolas da cidade de Belo Horizonte, a metodologia empregada na coleta buscou conceber a cidade de Belo Horizonte a partir de suas 9 regionais (Barreiro, Centro-sul, Leste, Nordeste, Noroeste, Norte, Oeste, Pampulha, Venda Nova). A fase de cadastro dos textos consistiu em sua digitalização e digitação, os textos foram escaneados em alta resolução, sendo armazenados em imagens de 3507x2480 pixels (mantida a taxa de 24 bits/pixel); a digitação dos textos foi feita manualmente por colaboradores do projeto. No intuito de alcançar a padronização dos textos, bem como um maior nível de similaridade entre o texto digitado e o original, um conjunto de 7 regras foi estabelecido (cf. CRISTÓFARO-SILVA et. al., 2006). Como resultado da metodologia adotada no Projeto e-Labore nas fases de coleta e digitalização dos dados, pode-se ter acesso às seguintes informações acerca das redações: número da redação, texto digitado (forma desviante e forma padrão), imagem digital da redação, nome do aluno, série, sexo, idade, nome da escola, tipo da escola (particular ou pública), número e data da coleta. A exceção da imagem digital da redação, todas as outras informações foram inseridas em um banco de dados em MySQL e organizadas por meio de tabelas. No atual estágio de desenvolvimento do trabalho, foco maior foi dado à informação contida no texto digitado. As palavras que continham desvios ortográficos foram separadas uma a uma e inseridas no banco de dados em MySQL na coluna formaDesviante, e sua forma padrão correspondente foi inserida na coluna formaPadrao. Buscou-se, a partir dos dados contidos nessas duas colunas, elaborar um algoritmo para classificar os desvios ortográficos. A classificação dos desvios foi feita tendo-se por base análises como as propostas por Scliar-Cabral (2003), Faraco (1997), Cagliari (1989) e Mollica (2003). Procurou-se atingir uma classificação geral dos desvios, observando-se os seguintes aspectos: troca, inserção ou apagamento de símbolos gráficos; troca, inserção ou apagamento de acento gráfico; troca entre letras maiúsculas e minúsculas; e junção ou separação de palavras. Cada um desses desvios constitui uma ou mais colunas na tabela principal do banco de dados, à qual se deu o nome dadosRedacoes. A estrutura atual dessa tabela possui 85659 tuplas, contendo 27 colunas, as quais estão descritas a seguir: id, codBarras, coleta, formaDesviante, formaPadrao, formaSonora, nomeAluno, serie, nomeEscola, tipoEscola, sexo, dataNascimento, dataColeta, desvioMaiFalt, desvioMaiSobr, desvioConsFalt, desvioConsSobr, desvioVogFalt, desvioVogSobr,

desvioAcentoFalt, desvioAcentoSobr, desvioEspacoFalt, desvioEspacoSobr, desvioHifenFalt, desvioHifenSobr, desvioTrocaCporC, desviotrocaCporV, desvioTrocaVporV, desvioTrocaVporC, e dadoVerificado. Todos as colunas que contêm informação acerca do tipo de desvio ortográfico, isto é, todas as colunas iniciadas por desvio, foram preenchidas automaticamente por meio de um algoritmo computacional implementado em PHP. Basicamente, o que o algoritmo fez foi percorrer cada um dos caracteres presentes em formaDesviante e comparar com aqueles presentes em formaPadrao, se os caracteres fossem idênticos, passava-se ao caractere seguinte, se fossem distintos, observava-se qual a diferença entre eles e o desvio era, então, marcado conforme a diferença observada. A taxa de êxito obtida com a execução do algoritmo foi de 84,0%, sendo classificadas 71978 das 85659 palavras do corpus. As palavras restantes foram classificadas manualmente. A estruturação do banco de dados em MySQL permite, através do cruzamento de informações, responder diversas questões de cunho linguístico ou para-linguístico. Observando-se a coluna formaDesviante, por exemplo, pode-se observar quais tipos de desvios ortográficos as crianças cometem. Cruzando-se os dados da coluna formaDesviante com os da formaSonora, por exemplo, é possível verificar quais os desvios ortográficos têm algum tipo de condicionamento fonológico ou não. De modo semelhante, pode-se obter respostas a perguntas de cunho para-linguístico: fazendo-se um cruzamento dos dados de desvio e a coluna tipoEscola é possível checar se há diferenças entre o número de desvios encontradas entre escolar públicas e particulares. Em suma, o corpus do Projeto e-Labore mostra-se como uma ferramenta de relevância para os estudos que abordem a aquisição da linguagem escrita, bem como sua relação com a fonologia. A organização do corpus em um banco de dados MySQL permite a realização de uma gama de opções de buscas, sendo possível e fácil o cruzamento das informações dentro do banco.

Contato: gustavocook@hotmail.com

VALIDAÇÃO DA TRANSCRIÇÃO E DA ANOTAÇÃO PROSÓDICA DE UM CORPUS ORAL

Heloísa Vale (PosLin/UFMG)
Maryualê Mittmann (PosLin/UFMG)
Priscila Osório Côrtes (IC / UFMG)

Este trabalho ilustra, quanto a corpora orais, uma metodologia inovadora para a validação da transcrição e a implementação da metodologia para a validação da segmentação prosódica. As validações são necessárias para conhecer o limite de confiabilidade estatística de qualquer estudo a partir destes corpora. Esta metodologia foi aplicada na compilação do corpus C-ORAL-BRASIL (Raso-Mello 2010), quinta ramificação do projeto internacional C-ORAL-ROM (Cresti-Moneglia 2005). O C-ORAL-BRASIL divide-se em uma metade formal e uma informal já pronta. A metade informal compõe-se de 139 textos e 210.000 palavras: 80% de contexto familiar/particular e 20% de contexto público; 1/3 de diálogos, 1/3 de conversações e 1/3 de monólogos. O corpus baseia-se essencialmente na diatopia mineira, respeita a variação diastrática, mas tem como principal objetivo a representação da variação diafásica, considerada a maior causa de variação na estrutura da fala. Com equipamentos de alta qualidade e sem fio foi possível gravar uma enorme variedade de situações comunicativas, inclusive em movimento. As transcrições são de base ortográfica (MacWhinney 2000), mas visam documentar muitas características da fala, possibilitando o estudo de fenômenos em curso de gramaticalização ou lexicalização: cliticização

dos pronomes sujeito, perda da morfologia verbal, perda da forma do verbo ser em estruturas clivadas, serialização verbal, aférese e muitos outros (Mello-Raso 2009). Durante as transcrições os textos foram segmentados prosodicamente em enunciados (quebras prosódicas percebidas como terminais, unidades com autonomia pragmática) e unidades tonais (quebras dentro do enunciado percebidas como não terminais) (Moneglia-Cresti 1997). Os transcritores/segmentadores passaram por uma fase de formação (Raso-Mittmann 2009). Quanto à segmentação prosódica, foram realizadas duas validações: uma antes de se iniciarem as transcrições (depois de vários treinamentos e avaliações) e uma depois que o corpus inteiro havia sido transcrito e revisado pela primeira vez, mas antes das revisões sucessivas. A primeira validação constitui uma implementação metodológica significativa: estabelece que o trabalho seja iniciado apenas quando há expertise que garanta um padrão qualitativo alto. Dessa maneira as fases de revisão podem efetivamente ocupar-se de refinamentos na transcrição/segmentação. A validação consistiu na obtenção de um grau de acordo, medido através de teste Kappa (Fleiss 1971), ≥ 0.8 para quebras terminais e ≥ 0.6 para as não terminais, entre 4 segmentadores. Os resultados gerais da primeira validação foram: 0.84 (quebras terminais) e 0.66 (não terminais), com algumas variações interessantes entre textos dialógicos e textos monológicos. Os resultados gerais da segunda validação foram 0.86 (terminais) e 0.78 (não terminais), com uma redução radical das diferenças entre textos dialógicos e monológicos. Um estudo qualitativo dos casos de desacordo revelou-se interessante tanto para identificar aspectos prosódicos do português brasileiro que induzem incertezas nos segmentadores, quanto para adquirir maior expertise para trabalhos futuros. A validação das transcrições constitui uma novidade no panorama metodológico da lingüística de corpus. O objetivo é quantificar o grau de confiabilidade das transcrições segundo duas perspectivas: a primeira, mais geral, consiste na quantificação da porcentagem de enunciados e de palavras com erros de transcrição. A segunda, essencial quando se usam critérios de transcrição não ortográficos, quantifica o grau de confiabilidade de cada critério de transcrição implementado. Aqui também a validação aconteceu em duas fases: antes da última revisão e depois da conclusão do corpus a ser publicado. Em cada fase extraiu-se uma amostra aleatória de 5% dos enunciados de cada texto e procurou-se por eventuais erros. A metodologia de busca foi a seguinte: dois transcritores verificaram as transcrições e, em caso de desacordo, recorriam a um terceiro transcritor. Esses casos foram raríssimos. Em princípio, consideramos satisfatória uma margem de erro não superior a 5% tanto para o total das palavras quanto para cada fenômeno considerado individualmente. Os resultados da primeira fase mostraram a presença de erros em 1,4% das palavras da amostra. A maioria destes consistiu de aplicação errada dos critérios não ortográficos, ou seja, o fenômeno foi codificado, mas de maneira diferente da convencionalizada. Trata-se do tipo de erro mais aceitável, porque recuperável mesmo posteriormente através da lista de palavras. Na análise dos erros por cada critério (35 diferentes fenômenos, por ex. você/ocê/cê ou para/pra/pa) verificamos 3% de erros (37 erros sobre 1165 ocorrências) sobre o total dos fenômenos, mas observamos: (i) que 9 dos fenômenos ocorreram em quantidades suficientes para termos uma amostra estatisticamente significativa (por ex. a oposição entre os pronomes pessoais tônicos ele/ela/eles/elas e a realização, geralmente clítica, como e'/ea/es/eas); (ii) que 9 dos fenômenos eram tão raros que a validação só poderia ser feita conferindo o corpus inteiro ou quase; esses fenômenos foram considerados estatisticamente não validáveis (por ex. a variação senhor/sior/sô); (iii) que 8 dos fenômenos não apresentavam um real interesse morfossintático (por ex. a variação hum hum/ham ham); portanto, independentemente de termos ou não alcançado uma base estatística significativa, foi encerrada a busca desses

fenômenos; e (iv) que 8 dos fenômenos apresentavam uma frequência ainda insuficiente para uma confiabilidade estatística, mas de grande interesse e com uma ocorrência tal que se poderia alcançar facilmente uma base estatisticamente significativa dobrando a amostra; nesse caso decidimos escolher, para cada texto, sempre aleatoriamente, mais 5% dos enunciados e, para alguns poucos fenômenos, mais 10%. Os resultados da primeira fase mostram o seguinte: para os fenômenos do grupo (i) na maioria dos casos não foram achados erros; o fenômeno que apresentou mais erros chegou a apenas 4% de erros (7 erros sobre um total de 172 ocorrências). Para os fenômenos do grupo (iv) os resultados foram muito diferentes dependendo do fenômeno: os erros variaram de um mínimo de 0 a um máximo de 14% (3 erros em 21 ocorrências). Uma nova validação, sempre sobre 5% dos enunciados, foi realizada após uma nova revisão das transcrições do corpus inteiro. Os resultados dessa segunda validação são os seguintes: verificamos 0,52% de erros sobre o total de palavras da amostra, que representam 1,01% de erros considerando-se apenas os fenômenos relacionados aos critérios não ortográficos. Essas metodologias de validação constituem um avanço na busca de confiabilidade dos corpora orais, mesmo sendo possível pensar em metodologias alternativas e até mais rigorosas.

Contato: helopv@terra.com.br

Referências:

CRESTI, E.; MONEGLIA, M. (eds.) (2005) C-Oral-Rom: Integrated Reference Corpora For Spoken Romance Languages. Amsterdam: John Benjamins.

FLEISS, J. L. (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382.

MACWHINNEY, B. J. (2000) The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum, 2 vol.

MELLO, H.; RASO, T. (2009) Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. Revista Portuguesa de Humanidades, v. 13, p. 301-325.

MONEGLIA, M.; CRESTI, E. (1997) Intonazione i criteri di trascrizione del parlato adulto e infantile. In: Bortolini, U.; Pizzuto, E. Il Progetto CHILDES Italia. Pisa: Del Cerro, pp. 57-90.

RASO, T.; MELLO, H. (2010) The C-ORAL-BRASIL corpus. In: MONEGLIA, M.; PANUNZI, A. (orgs.) Bootstrapping Information from Corpora in a Cross Linguistic Perspective. Firenze: Firenze University Press. p. 193-213.

RASO, T.; MITTMANN, M. M. (2009) Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. Revista de Estudos da Linguagem, 17, 73-91.

NANOCIÊNCIA/NANOTECNOLOGIA E BIOCOMBUSTÍVEIS VISTOS PELO MODELO SILEX: ANÁLISE MORFOLEXICAL DE TERMINOLOGIAS

Joel Sossai Coleti (UFSCar/Fapesp)

A estrutura de dois repertórios terminológicos (Nanociência/Nanotecnologia e Biocombustíveis) é descrita, nesta pesquisa, no nível da estrutura interna dos seus termos constitutivos, verificando-se os principais processos de construção dos termos (tipos de derivação presentes, tipos de composição – morfológica, morfossintática, lexicalização de sintagmas, etc.). A partir dessa descrição morfológica, organiza-se uma base de dados de maneira com vistas à posterior implementação computacional e disponibilização on-line dos dados obtidos. Utiliza-se o modelo SILEX, modelo de morfologia construcional concebido por Danielle Corbin e desenvolvido pela autora e pelos membros do centro de investigação SILEX (Syntaxe, Interprétacion, et LEXique), sediado na Universidade de Lile III (França), hoje denominado STL (Savoirs Textes Langage). O modelo adotado tem sido aplicado com êxito à descrição do léxico construído de várias línguas românicas, como francês e grego moderno, no tocante a língua portuguesa a pesquisadora Graça Maria Rio-Torto tem sido sua grande impulsionadora. A escolha desse modelo teórico não é inocente: o objetivo do modelo SILEX é construir uma teoria sincrônica do léxico capaz de atribuir uma estrutura e uma interpretação adequadas às palavras construídas, atestadas ou não nos dicionários, de modo a caracterizar a natureza da "gramaticalidade lexical" e de determinar as restrições das regras de formação de palavras. A base teórica do modelo é construída em linhas gerais em D. Corbin 1987 ,1991 e 1999. Segundo a autora, a originalidade do modelo que propõe reside no refinamento da tipologia das associações entre forma e significado características das palavras construídas e na mudança da habitual ordem de prioridades presente nos trabalhos de morfologia derivacional: em vez de propor análises baseadas na evidência do léxico observável, propõe uma análise baseada na estratificação e na reconstrução do léxico descritível; em vez de dar prioridade à análise morfológica sobre a análise semântica, propõe uma análise que associa forma e significado. Dessa forma, torna-se um modelo de morfologia capaz de lidar com o tratamento da semântica e da referência das palavras construídas. Além disso, o modelo SILEX permite a análise dos processos que levam determinadas unidades – provenientes da língua corrente ou dos vocabulários de outros domínios especializados – a tornarem-se aptas para denominar conceitos próprios de domínios especializados. O modelo SILEX assume-se como um modelo associativo e estratificado. Por 'modelo associativo' entende-se aquele cujas Regras de Construção de Palavras (RCPs) permitem construir conjuntamente a estrutura morfológica e a interpretação semântica das palavras construídas; É um 'modelo estratificado' porque apresenta um componente lexical da gramática composto por vários níveis, ao longo dos quais se vai construindo o significado das palavras construídas. Este modelo oferece, também, a aparelhagem teórica necessária para dar conta não apenas da estrutura morfológica dos termos, mas também da polissemia e da polirreferência das unidades que integram os vocabulários em estudo. Nesta oportunidade serão apresentadas análises das terminologias de Nanociência/Nanotecnologia e Biocombustíveis, repertórios terminológicos extraídos de corpora. Vale ressaltar que a compilação dos corpora baseou-se nos requisitos da Linguística de Corpus, tais como: autenticidade, representatividade, balanceamento, amostragem, diversidade e tamanho. Com os corpora prontos, avançou-se para a fase de extração semiautomática dos candidatos a termos, utilizando

para essa tarefa o software Ngram Statistics Package (NSP). Na sequência a listas geradas pelo extrator foram validadas manualmente. (Apoio: FAPESP)

Contato: joelscoleti@gmail.com

METÁFORA GRAMATICAL EM UM CORPUS DE APRENDIZES

Lucia Oliveira (PUC-Rio)
Violeta Quental (PUC-Rio)
Rubiane Valerio (PUC-Rio)
Adriana Nobrega (PUC-Rio)
Maria Cristina Monteiro (PUC-Rio)
Ana Elisa Vianna (PUC-Rio)
Vera Lucia Selvatici (PUC-Rio)

O objetivo deste trabalho é apresentar os resultados de um estudo de corpus sobre a produção textual em escolas públicas e particulares do Rio de Janeiro, tendo como base teórica a perspectiva sistêmico-funcional. Este estudo faz parte do projeto 'Escrita e inclusão social: o corpus e a metáfora gramatical no Ensino Médio' (FAPERJ, 2008), em que as nominalizações foram identificadas no corpus com a ajuda de software específico (Scott, 2004) e analisadas qualitativa e quantitativamente visando-se estudar sua frequência e seu uso na estrutura temática das orações, bem como a incidência de argumentos (participantes ou circunstâncias) que as acompanham. Um corpus de redações de alunos da 1ª à 3ª série do ensino médio (N= 651 textos) foi coletado, organizado e codificado visando ao estudo da metáfora gramatical (Halliday,1994), que implica, dentre outros aspectos, a transformação de ideias mais concretas em mais abstratas, através do uso de nominalizações em lugar de processos verbais, podendo constituir-se em dificuldade específica para o domínio da escrita (Simon-Vanderbergen et al, 2003, Webster e Halliday, 2009). Os resultados da pesquisa indicam o uso reduzido da metáfora gramatical no corpus de aprendizes, havendo variação na frequência e proficiência na utilização desse recurso lingüístico nos diferentes grupos de textos analisados. Visando-se discutir a aplicação da teoria sistêmico-funcional a textos produzidos em português e o uso da metodologia de corpus em textos de alunos em língua portuguesa, serão discutidas questões relativas à descrição e aplicação de categorias sistêmicas ligadas à metáfora gramatical, bem como problemas quanto à construção de corpus de aprendizes, do ponto de vista de sua análise por ferramentas computacionais (Bick, 2000; Paumier, 2000).

Contato: luciapo@openlink.com.br

CONSTRUÇÃO E ANÁLISE DE CORPUS DE LÍNGUA FALADA EM ESTUDOS DE INTERPRETAÇÃO E TERMINOLOGIA – O CASO DO CORPUS CAFÉ

Luciana Latarini Ginezi (USP)

O presente trabalho baseia-se em pesquisa desenvolvida na área de Linguística de Corpus, aliada aos Estudos de Interpretação e Terminologia. Os Estudos de Interpretação estão inseridos nos Estudos de Tradução (Williams & Chesterman, 2002), pois reúne características semelhantes em termos de função social, mediação cultural, dentre outras. Sabemos que a Interpretação, assim como a Tradução, prevê a criação de glossários como parte do processo de trabalho (Gile, 1995). Para seu trabalho, o intérprete constrói um glossário específico, contendo termos e fraseologias, bem como acrônimos, siglas e outras informações necessárias para a oralidade. A oralidade, por sua vez, não permite que o intérprete possa pesquisar termos que não constem no glossário, devido ao imediatismo implícito em sua concepção. Em Terminologia, as variantes estão presentes como parte da linguagem de especialidade, principalmente quando se trata da língua falada, objeto de trabalho do intérprete. Assim, para que uma interpretação de área técnica seja feita, percebe-se a importância do conhecimento das variantes daquela língua de especialidade. Por isso, este trabalho buscou identificar variantes terminológicas na língua falada, da área de especialidade do Café, para observar as diferenças e semelhanças entre língua escrita e língua falada no âmbito da preparação do intérprete para seu trabalho. Partindo da hipótese de que há mais variantes na língua falada, conduziu-se a pesquisa a fim de observar tal ocorrência, bem como analisar os resultados produzidos do levantamento terminológico. A área do Café foi escolhida devido ao trabalho da pesquisadora em interpretação estar relacionado ao assunto, bem como devido à inexistência de produtos terminológicos bilíngues de apoio a tradutores e intérpretes na área. Sendo o Brasil o maior produtor mundial de café tipo arábica, a necessidade de se produzir um material de consulta confiável é primordial. Para a coleta de dados orais, foram conduzidas entrevistas com especialistas da área, trabalhadores rurais, administradores de fazendas, agrônomos, executivos de empresas de vendas de máquinas agrícolas, ou seja, profissionais de categorias diversas, representantes das funções possíveis da área do Café, desde que sobre os assuntos colheita e processamento, selecionados por sua relevância. Também foram gravadas interpretações realizadas no modo Consecutivo. As línguas de trabalho foram inglês e português, com falantes brasileiros da região Sul de Minas Gerais ou da Alta Mogiana Paulista, e estrangeiros falantes de inglês de diversos países, dentre eles Austrália, Canadá, Estados Unidos da América, Etiópia, Quênia etc. A seleção dos falantes foi feita devido às interpretações reais que aconteciam nas regiões brasileiras citadas. Além disso, em uma viagem à Etiópia, berço produtor de café mundial, a pesquisadora pode realizar o trabalho de gravação de conversações e entrevistas durante visitas a fazendas de café etíopes e ao evento Eastern African Coffee Association Conference & Exhibition, em Addis Ababa, onde houve a participação de profissionais do café de todo o mundo e cuja língua oficial era inglês. Após a coleta de dados, várias formas de transcrição foram efetuadas, com programas computacionais e aparelhos de gravação sonora diferentes, tentando encontrar uma maneira menos árdua de finalizar essa etapa do processo de construção do corpus. Ao finalizar a construção dos corpora (monolíngues – entrevistas e conversação; bilíngues – interpretação), utilizamos a ferramenta Wordsmith Tools para levantamento e análise terminológica, através das keywords e das concordâncias produzidas. Com o objetivo de comparar os resultados terminológicos produzidos entre modalidade escrita e oral

da língua, também foram construídos corpora em língua escrita, em inglês e português, a partir de textos informativos coletados na web. Os resultados da comparação demonstram o quanto a língua falada possui mais variantes em relação à língua escrita. A partir deles, produzimos um glossário terminológico do Café, utilizando ambos os corpora falados e escritos, com a inclusão de variantes terminológicas provenientes da oralidade.

Contato: lucianaginezi@uol.com.br

FICTIVIDADE, DISCURSO E LINGUÍSTICA DE CORPUS: O CASO DE AUTOCITAÇÃO FICTIVA

Luiz Fernando Matos Rocha (UFJF)

Os estudos sobre fictividade (TALMY, 1996, 2000; LANGACKER 1991, 1999, 2008; PASCUAL, 2006; BRANDT, 2010) dão conta de que certas expressões linguísticas estão apenas indiretamente vinculadas a seus referentes pretendidos e que cenários não-verídicos são frequentemente apresentados pelos usuários das línguas com o propósito de obter acesso mental aos cenários efetivos. No exemplo, "A cerca vai do platô até o vale", uma parte de nossa cognição pode muito bem perceber a imagem de algo em movimento, percorrendo o caminho que vai do platô ao vale. Não obstante, outra parte de nossa cognição pode avaliar essa imagem como irreal, preservando a concepção de que nada na cena está se movimentando na realidade. Considerando esse tipo de conflito cognitivo interno, a imagem avaliada como irreal é fictiva. Agregando frames cognitivos e interacionais, o fenômeno da autocitação fictiva em especial é um tipo discursivo de fictividade por meio do qual seus conceptualizadores impõem uma perspectiva subjetificante e avaliativa ao discurso direto em primeira pessoa, diferentemente de sua contraparte factiva. Isso é provocado sobretudo pelo uso incongruente entre a forma canônica de reportar a fala ou o pensamento próprios e o sentido de verbos dicendi, como "dizer" e "falar", que passam a assumir status exclusivamente epistêmico (e.g. Eu disse (pensei): "Ai, meu Deus!"). Assim, por meio de um cenário não-verídico de reportação discursiva, o agente locutório remete-se a um cenário prévio e suposto de fala, com propósito efetivo de permitir acesso mental ao cenário verídico de pensamento. O percurso histórico-metodológico dos estudos sobre fictividade é similar ao da Linguística Cognitiva como um todo. Tem início com trabalhos que se baseiam puramente na intuição dos linguistas, que desenvolvem construtos epistemológicos induzidos a partir de ilustrações imagísticas e linguísticas, inventadas ou artificiais — embora plausíveis —, para a postulação de realidades psicológicas e cognitivas. Por sua vez, o objetivo central deste trabalho é a descrição e análise da autocitação fictiva e sua co-extensão factiva em corpora orais de Português Europeu (PE) e Brasileiro (PB), a partir da construção semiaberta (EU) disse/falei X-oracional, despida de quaisquer sintagmas direcionais (GOLDBERG, 1995) ou zonas ativas (LANGACKER, 1991), que apontariam inequivocamente para sua interpretação factiva. Utilizam-se como base de dados o corpus C-ORAL-ROM Português (NASCIMENTO, GONÇALVES, VELOSO, ANTUNES, BARRETO E AMARO, 2005) e o corpus C-ORAL Brasil (RASO e MELLO, 2010), de arquitetura básica similar, bem como os corpora CINTIL, NURC e o banco de dados do reality show Big Brother Brasil (2002); todos submetidos às ferramentas eletrônicas TextSTAT ou Contextes. Em geral, os resultados apontam para contrastes conceptuais, diatópicos e diafásicos significativos entre usos de "disse" e "falei" nas variedades nacionais, uma vez que o verbo "falar" não costuma

ser usado para introduzir espaço mental de discurso reportado em PE e que certos frames interacionais são colostrucionalmente mais propícios ao surgimento de autocitação fictiva, como o reality show. No entanto, a fictividade afeta discursivamente a autocitação em ambas as variedades nacionais do português, mapeada por pistas que incluem reportação monológica, subjetificação, co-texto epistêmico, escaneamento mental, metáfora PENSAMENTO É FALA, atos de falas como promessa, planejamento e apreciação, reivindicação de face de amizade e incongruência dêitica. Tais sinais formam um conjunto de tendências semântico-pragmáticas extraído do exame caso-a-caso de interações reais, fazendo convergir frames interacionais e cognitivos que sustentam a natureza multidimensional do fenômeno, basicamente dividida em dimensões epistêmicas e pragmáticas. Isso contribui para um entendimento inédito acerca da fictividade que, segundo Talmy (2011), diz respeito apenas a conflitos cognitivos internos entre modos discrepantes (fictivos e factivos) de se perceber ou conceber o mesmo objeto. Por outro lado, se levamos em consideração a hipótese colostrucional de Gries e Stefanowitsch (2003, 2004, 2008), que envolve essencialmente a identificação da força de associação entre uma determinada construção e um item lexical, e a tratamos em termos discursivos, chegamos à conclusão de que um frame cognitivo fictivo é evocado quando um frame interacional fictivo também o é.

Contato: luiz.rocha@ufjf.edu.br

COMO UM ESTUDO BASEADO EM CORPORA PARALELOS PODE APRIMORAR A ELABORAÇÃO DE VERBETES BILÍNGUES?

Marion Celli (USP/Fapesp)
Adriana Zavaglia (USP)

A linguística de corpus vem se destacando, nas últimas décadas, tanto em pesquisas monolíngues quanto multilíngues, sendo, desde a década de 1960, muito utilizada em estudos sobre a elaboração de dicionários. Pesquisas baseadas em corpora mostram-se assim cada vez mais fundamentais para a lexicografia moderna, em que grandes quantidades de texto são analisadas num baixo período de tempo e com maior credibilidade. Apesar da evolução de centros de pesquisa nessa área – principalmente na Europa –, ainda são poucos, no Brasil, os estudos voltados para o campo da lexicografia bilíngue referentes à língua geral, mais especificamente, às palavras ditas gramaticais. Em um levantamento anterior sobre a descrição de marcas adversativas do português brasileiro em dicionários bilíngues (I) português-francês e (II) português-inglês, por exemplo, foi observada não apenas a ausência de contextualizações e especificidades semânticas, sintáticas ou pragmáticas nos verbetes, como a baixa variabilidade tradutória oferecida para cada unidade observada, a saber: mas, porém, contudo, todavia e entretanto. Sem definições e exemplificações, como um aprendiz, um tradutor, ou qualquer outro interessado na língua-alvo saberia identificar, dentre as opções dadas para todavia, por exemplo, a melhor alternativa em inglês ou francês? Nevertheless ou notwithstanding? Ou será however? Néanmoins ou cependant? Talvez toutefois? A partir dos verbetes analisados, um consulente certamente teria dificuldades na escolha do correspondente mais coerente para o contexto exigido na língua-fonte, já que, nos dez dicionários português-francês e sete português-inglês consultados, nenhum apresentou exemplos de uso das opções oferecidas em francês ou inglês.

Além disso, sem uma descrição de suas diferenças, poderia inferir que os correspondentes são substituíveis entre si em quaisquer contextos – fato que pesquisas baseadas em corpus não corroboram. Apesar de em alguns casos – Burtin-Vinholes (1993), para o francês, e Michaelis (2000), para o inglês – termos certa variedade tradutológica, incluindo agrupamentos ou expressões (mas também – mais encore; nem mas nem meio mas – but me no buts), percebe-se que não há reflexão sobre o uso de cada uma das formas apresentadas. No entanto, são muitas as particularidades de pontuação, posicionamento e registro entre uma língua e outra que unidades gramaticais, tal como as lexicais, também podem trazer em sua definição lexicográfica. Ademais, é válido lembrar que, intrinsecamente ligada ao ato tradutório, está a interface entre língua e cultura. Mesmo quando se traduz marcas gramaticais é preciso considerar que cada sistema cultural não apenas condiciona a visão de mundo do homem como também possui a sua própria lógica (Laraia, 2008). Observa-se, dessa maneira, a intrigante tarefa do pesquisador em encontrar os traços linguísticos que revelam tais visões desencontradas de mundo na relação entre o texto-fonte e o texto-alvo, fatores essenciais para elaboração de verbetes bilíngues. A busca por correspondentes tradutórios gramaticais não pode, desse modo, isolar-se do contexto situacional empregado, que caracteriza, na primeira ou na segunda língua, variações de ordem cultural que devem ser levadas em consideração. É nesse contexto que a associação de corpora paralelos e lexicografia bilíngüe torna-se interessante. Tendo em vista tal problemática, pretendemos, a partir da apresentação dos verbetes bilíngues (I) português-francês e (II) português-inglês elaborados para mas, porém, contudo, todavia e entretanto, demonstrar a importância de estudos baseados em corpora para a construção de verbetes bilíngues definidos e contextualizados. Após um estudo monolíngue das unidades no português brasileiro e da subseqüente análise tradutológica das marcas, em francês e em inglês, foram elaborados cinco verbetes português-francês e cinco português-inglês, sendo dois verbetes para cada marca adversativa estudada. O corpus foi constituído de cinco obras da literatura brasileira do século XX (Sagarana, de Guimarães Rosa; A República dos Sonhos, de Nélida Piñon; A paixão segundo G.H., de Clarice Lispector; Onde andará Dulce Veiga, de Caio Fernando Abreu; e Benjamim, de Chico Buarque) e suas respectivas traduções para o francês e o inglês (1.582.305 palavras) tendo, como suporte computacional, os programas WordSmith Tools e ParaConc. Com o auxílio do WordSmith Tools, foi possível, inicialmente, analisar o comportamento enunciativo (Culioli, 1990) das unidades em contexto monolíngue para, em seguida, observar, a partir do programa ParaConc, as traduções para cada ocorrência levantada. Considerando o corpus de estudo, pôde-se concluir que, devido às relações interlinguísticas e ao caráter transcategorial das palavras (Culioli, 1990), mas, por exemplo, não terá como correspondente, em francês ou em inglês, necessariamente uma conjunção coordenativa adversativa, assim habitualmente definida. Graças a fatores semânticos, contextuais, estruturais, estilísticos, etc., pode-se ter, como possibilidades tradutórias, plutôt (fr.) e what about (ing.), por exemplo. Além disso, há casos em que mas é omitido, havendo, assim, uma supressão de parte do texto fonte. É possível afirmar, desse modo, que, a partir do corpus paralelo observado, não é somente considerável a variação tradutológica de mas, porém, contudo, entretanto e todavia – tanto em francês quanto em inglês – como também cada tradução apresenta particularidades semânticas, sintáticas e pragmáticas interessantes. É importante, dessa forma, ressaltar a necessidade de definição e contextualização em verbetes bilíngues, já que os modelos tradicionais não levam em conta aspectos sintático-pragmáticos como os levantados neste trabalho.

Contato: marion.celli@gmail.com

A Construção Superlativa Sintética de Estados Absolutos com o Sufixo –íssimo: um estudo baseado em corpus

Patrícia Miranda Machado UFJF)
Neusa Salim Miranda (UFJF)

Assumindo a perspectiva sociocognitiva e construcionista configurada pela Linguística Cognitiva (LAKOFF, 1987, 1993; JOHNSON, 1987; LAKOFF & JOHNSON, 1980[2002], 1999; SALOMÃO, 1999, 2009a, MIRANDA, 1999; CROFT & CRUSE, 2004; SILVA, 1997) e pelos Modelos de Uso da Gramática das Construções (GOLDBERG, 1995, 2006; CROFT & CRUSE, 2004; SALOMÃO, 2009b; MIRANDA, 2008b, TRAUGOTT, 1995), o presente trabalho busca investigar um dos nódulos da rede de Construções superlativas do Português aqui nomeado como Construção Superlativa Sintética de Estados Absolutos (CSSEA) do tipo "Olha no momento eu tou eh desempregadissima da silva!"; "Tenho vinte e poucos anos, sou casadíssima e muito feliz!"; "Gravidíssima, Negra Li abre o closet e mostra o que veste". Trata-se de uma construção morfológica formada a partir da integração de um núcleo que remete a um estado absoluto não-graduável (desempregada, casada, grávida) com um operador de escala superlativa (- íssimo/a). O resultado são types como desempregadíssima, casadíssima, gravidíssima, formadíssima, dentre outros. A natureza dos estudos sociocognitivos e construcionistas sugere um recorte epistemológico que confere ao USO papel fundamental na emergência da Gramática e do Léxico de uma língua (MIRANDA, 2008b, p. 4). Nesse sentido, o conhecimento linguístico de um falante é visto como uma rede de símbolos erguidos na cultura através do uso (MIRANDA, op. cit.) e apreender a real natureza desse conhecimento só é possível se o observarmos dentro das molduras que configuram o discurso real. Tais supostos implicam um sólido compromisso com a empiria e direcionam nossas escolhas metodológicas para a análise de corpus o que implica, em uma visão probabilística da linguagem, a utilização de uma base volumosa de dados, capazes de fornecer indicadores acerca do uso da construção (frequência de types e tokens) em seu habitat discursivo real. Dentro dessa abordagem, procedeu-se à constituição de um corpus específico da construção baseado em dados reais e espontâneos de uso linguístico, através do concordanciador eletrônico Web Concordancer beta - http://webascorpus.org/searchwac.html. O uso de tal ferramenta deveu-se ao fato de a construção em foco, dada a sua marca de informalidade, apresentar um baixíssimo nível de ocorrência nos corpora tratados testados (Corpus do Português e VISL). Nos termos apresentados, nosso corpus se configura a partir de um universo total de 8.189.656 palavras em que foram investigados 30 types e registrados 1.757 tokens. A configuração da CSSEA aponta para o fenômeno do desencontro/ mismatch (FRANCIS & MICHAELIS, 2000; TRAUGOTT, 2007; TRAUGOTT, 2006), uma vez que evidencia incongruências entre as propriedades semântico-formais das unidades que integram este padrão - o afixo superlativo -íssimo e o item lexical por ele graduado. Essa incongruência da construção é abordada a partir das relações polares de contrário e contradição, descritas por Israel (2004). Assumindo a hipótese de que a CSSEA é uma construção, propomos, como tarefa analítica principal, a descrição de seus polos formal e semântico-pragmático. Quanto aos aspectos formais, encontramos três padrões construcionais da CSSEA definidos por núcleos Adjetivos, Substantivos e Adverbiais. O padrão formal majoritário da CSSEA (76,5%) é constituído por radicais deverbais participiais (casado, namorado, eleito, comprado...). No que se refere à Semântica da CSSEA, nossas análises desvelaram os frames a que

se vinculam os seus types de modo a recobrir o novo perfilamento dado a tais cenas. Assim, postulamos a presença de dois tipos de perfilamento nos contextos discursivos da CSSEA: i. Perfilamento por Traços; ii. Perfilamento por Contradição. Para além de identificar um estado, ambos os perfilamentos que envolvem a CSSEA atribuem-lhe propriedades consensuais, idealizadas, apontando em direção à noção de protótipo. A configuração do habitat discursivo da CSSEA, caracterizado pela informalidade, delineou-se através de três categorias de análise em relação ao texto fonte: (i) temática principal, (ii) gêneros discursivos e (iii) público alvo. Os resultados obtidos em relação a tais categorias, mostraram que os usuários da língua recorrem, de modo geral, à CSSEA para falar, de maneira descontraída, do cotidiano, daquilo que os diverte, de suas vidas, das vidas de outras pessoas, dos artistas etc., através de posts de blogs, comentários, notícias, anúncios, sem restringir, contudo, o gênero do público alvo. As análises empreendidas, neste estudo, ancoradas em padrões de frequência, puderam consolidar a hipótese inicial de que a CSSEA se constitui como um padrão construcional de uso específico dentro da rede de Construções Superlativas do Português. A frequência de types (30) e de tokens (1.757) atestaram, respectivamente, a produtividade significativa da construção e o processo de convencionalização de alguns de seus types no Português do Brasil (como aprovadíssimo, recomendadíssimo, solteiríssimo, candidatíssimo e confirmadíssimo) em ambientes discursivos marcados pela informalidade. Este estudo, buscando descrever um nódulo morfológico da rede de Construções Superlativas do Português, ilustra a virada metodológica promovida pelos estudos sociocognitivos e construcionistas da gramática e do léxico. De igual modo, desvela a relevância posta no uso e na diversidade linguística.

Contato: patmmachado@gmail.com

COMO SE DIZ "FAZER UM GOL" EM INGLÊS? FRASEOLOGIA DO FUTEBOL

Sabrina Matuda (USP)

Esta apresentação, resultado de nossa pesquisa de mestrado desenvolvida na Universidade de São Paulo, mostra um estudo aprofundado da terminologia do futebol em português e inglês por meio do estabelecimento de equivalentes fraseológicos. O futebol é o esporte mais praticado no Brasil e no mundo. É reconhecido mundialmente enquanto competição, manifestação cultural e até mesmo como um mercado na ordem econômica. É fato que as relações futebolísticas entre o Brasil e os países da Europa crescem cada vez mais (CRUZ, 2005); seja pelo intercâmbio de contratos de jogadores e técnicos; pelos direitos de transmissão de campeonatos; pelo patrocínio de jogadores por grandes marcas ou por qualquer outra negociação que envolva um produto relacionado ao futebol. Para que todas essas relações se materializem, estabelecemos uma comunicação que, na grande maioria das vezes, se dá no uso da língua inglesa. No entanto, cada nação tem a sua maneira de jogar, torcer e narrar, maneira, esta, expressa por meio da nossa língua materna. O problema surge quando queremos expressar essas particularidades em uma língua estrangeira. Realizamos um levantamento das obras terminográficas sobre futebol bilíngues (português-inglês) existentes no mercado e constatamos a escassez de materiais elaborados com rigor metodológico e a falta de padronização no estabelecimento de equivalentes terminológicos. Ademais, verificamos que os dicionários disponibilizam, na maioria das vezes, somente equivalentes para os termos, raramente incluindo um colocado ou a frase em que o termo ocorre.

Outra característica comum é a ausência de exemplos, ou seja, ao usuário é apresentado o termo equivalente, mas sem informações, quer das palavras que o acompanham, quer do contexto no qual é utilizado. Dada a escassez de obras terminográficas, a importância da eficácia na comunicação especializada e a falta de padronização na tradução das fraseologias, o objetivo principal de nossa pesquisa foi fazer um estudo aprofundado da terminologia do futebol por meio do estabelecimento de equivalentes fraseológicos. Os objetivos específicos do estudo das unidades fraseológicas especializadas (UFEs) foram: 1) a partir da compilação de um corpus comparável português-inglês de futebol, identificar as UFEs mais frequentes no corpus de português e seus equivalentes no corpus de inglês; 2) criar um modelo de glossário bilíngue português-inglês de fraseologias do futebol; 3) elaborar um verbete do glossário tendo como base o modelo proposto. A fim de atingir os objetivos propostos, compilamos um corpus comparável português-inglês de futebol composto por, aproximadamente, um milhão de palavras em cada língua: 917.073 em português e 1.002.897 em inglês. Cada corpus é dividido em quatro subcorpora: regras do jogo, textos jornalísticos sobre resultados de partidas, narrações minuto a minuto e "transmissões sociais". A análise do corpus foi realizada de modo semi-automático, utilizando o etiquetador Tree-Tagger para fazer a etiquetagem morfossintática dos textos e o programa WordSmith Tools para explorar o corpus. Para conduzir a análise, utilizamos quatro abordagens teóricas: a Linguística de Corpus (LC), a Terminologia Textual, a Tradução Técnica como ato comunicativo sujeito a condicionantes culturais e o conceito forma-representação. A compilação e a exploração dos corpora foram realizadas utilizando os conceitos subjacentes à LC (BERBER SARDINHA, 2004; TOGNINI-BONELLI, 2001; PEARSON, 1998; SINCLAIR, 2004). A Terminologia Textual (FINATTO; KRIEGER, 2004) nos deu o aporte teórico para o estabelecimento dos critérios de identificação bem como para a definição das UFEs. As noções de equivalência de Azenha Jr. (1999) e Tagnin (2007), utilizadas na tradução técnica, contribuíram de forma significativa para nossa concepção de equivalente fraseológico. O conceito forma-representação de Toledo (2002) foi fundamental para entendermos algumas discrepâncias na linguagem utilizada para falar de futebol em inglês e português. Ressaltamos que, para esta apresentação, trabalhamos somente com as UFEs do termo gol, palavra-chave mais frequente do corpus de português. Em síntese, a metodologia adotada para o levantamento das UFEs foi: 1) gerar palavras-chave (keywords) do corpus em português para extrair os candidatos a termo; 2) selecionar o item mais frequente da lista, o termo "gol"; 3) gerar linhas de concordância de goal e examinar os clusters para extrair os candidatos a fraseologias; 3) unificar os clusters que fazem parte de uma mesma unidade de sentido e gerar linhas de concordância para os clusters em busca de padrões maiores para validar as UFEs; 4) gerar linhas de concordância da tradução prima facie de goal e de algumas etiquetas gramaticais para estabelecer os equivalentes fraseológicos. Identificamos 58 UFEs no corpus de português e 102 possíveis equivalentes no corpus de inglês e criamos um modelo de glossário bilíngüe português-inglês de fraseologias do futebol com três níveis para os verbetes. No primeiro nível, apresentamos o termo isolado, por exemplo, gol; no segundo, apresentamos as UFEs como, por exemplo, sair do gol; e no terceiro nível, elencamos as UFEs que podem fazer parte da UFE apresentada no segundo nível: [goleiro] SAIR do gol e ficar com a bola. Com base no modelo proposto, elaboramos um verbete para o termo gol com as 58 UFEs identificadas no corpus de português. Todas as entradas possuem equivalentes em inglês e exemplos; quando necessário, apresentamos formas sinônimas, remissivas e comentários. As reflexões do estudo nos mostraram que a Terminologia não é uma atividade prescritiva, na qual os termos devem ser normatizados a fim de garantir a eficácia de uma comunicação especializada. Ao

contrário, os fatores culturais, o contexto e a situação, a finalidade de uso e o estilo de futebol de cada país influenciam de forma direta o funcionamento das terminologias; por esse motivo, o fazer terminológico, principalmente o bilíngue, deve considerar todos esses elementos na compilação de obras terminográficas.

Contato: sa_brina1@yahoo.com.br

O VERBO 'GET' E SUAS MIL E UMA UTILIDADES: COMPARA E GOOGLE TRADUTOR NO ENSINO E APRENDIZAGEM DE INGLÊS COMO LÍNGUA ESTRANGEIRA

Tarsila Rubin Battistella (Unisinos)
Dra. Isa Mara da Rosa Alves (Unisinos)
Marília dos Santos Lima (Unisinos)

Refletir sobre as contribuições da Linguística de Corpus (LC) para o ensino de língua estrangeira tem sido tema de relevância reconhecida entre pesquisadores (ASTON 1997; BERBER SARDINHA 2006; JOHNS 1991; HUNSTON 2005). Entre os professores de línguas, no entanto, não vemos essa aproximação com a mesma frequência, mas entendemos que ela é altamente salutar. Tendo em vista tal cenário, apresentamos nesta pesquisa uma proposta de trabalho que se propõe a explorar corpus como material de ensino (cf. FLIGELSTONE, 1993). Este trabalho descreve e reflete sobre uma atividade pedagógica para o ensino do verbo get para alunos de aviação através das ferramentas COMPARA e Google Tradutor. Concordamos com Berber Sardinha (2006) quando ele diz que o trabalho com corpus pode auxiliar o aluno e o próprio processo de aprendizagem, tendo em vista que possibilita uma experiência mais consciente com o léxico, possibilitando a ele abandonar a concepção de que o vocabulário de uma língua representa um simples "conjunto de palavras isoladas que se juntam por meio de regras gramaticais" (p.149). O aluno começa a entender então que as palavras se aproximam por atração mútua, formando o "tecido da linguagem" (BERBER SARDINHA, 2006, p.149). Aproximar a LC e o ensino de inglês como língua estrangeira representa uma atitude coerente com a perspectiva sociocultural de ensino, que concebe o aprendizado e o desenvolvimento dos aprendizes de acordo com a participação deles em atividades socioculturais em suas comunidades (LANTOLF, 2000). A teoria tem origem a partir de Vygotsky (2003) e relaciona-se à mente, que reconhece o papel central das relações sociais e artefatos construídos culturalmente na organização das formas de pensamento. Para tal teoria, a aprendizagem ocorre no momento em que o aprendiz interage através da linguagem e, depois, internaliza o conhecimento que foi construído a partir das relações interativas, apropriando-se dele (LANTOLF, 2000). Nesse sentido, a aprendizagem é de certa forma, um processo de internalização. O aprendiz converte as ações externas às quais ele é exposto, em funções psicológicas superiores, consolidando assim o aprendizado. Com base nos apontamentos feitos acima, elaboramos um experimento que tem como objetivo auxiliar na aprendizagem do inglês como língua estrangeira de comissários de voo. Para tal, criamos estratégias para a observação de múltiplos sentidos do verbo get, através dos recursos COMPARA e do Google Tradutor. A escolha pelo trabalho com get deve-se ao fato de o verbo representar um dos principais desafios para os alunos, nesse caso, comissários, em razão dos múltiplos sentidos e padrões de ocorrência

convencionalmente associados a ele. Num contexto geral, para aprender os múltiplos sentidos frequente associados a tal verbo, é necessário que se entenda e se pratique o uso do verbo de uma forma consciente, dentro de contextos reais de uso. Sabe-se que decorar tais sentidos não garante conhecimento suficiente para fazer uso do verbo em diferentes situações de comunicação. Tal fato evidencia que para que o aluno entenda os usos de 'get' ele precisa observar o contexto em que se insere e as ferramentas de corpus possibilitam isso. O corpus paralelo permite buscar por uma palavra, em português ou em inglês, e visualizar, lado a lado, diferentes contextos frasais em que tal item ocorre tanto em português quanto em inglês, graças a um trabalho de tradução humana. O COMPARA foi utilizado no experimento por possibilitar a observação das ocorrências de get e de seus equivalentes de tradução em português em concordâncias paralelamente alinhadas. O Google Tradutor foi escolhido como ferramenta pedagógica complementar por ser geralmente o recurso consultado pelos alunos como dicionário e tradutor, que permite verificar a pronúncia, a tradução e a versão. Para a realização do experimento, as seguintes etapas foram realizadas: (1) pesquisa do significado do verbo no Google Tradutor e (2) pesquisa ultra-avançada do verbo 'get' no COMPARA. Na ferramenta COMPARA tornou-se possível realizar uma pesquisa do inglês para o português com o verbo 'get' de concordância em um contexto e foi possível também selecionar a variedade de português brasileiro e de inglês americano e britânico. Após a pesquisa no COMPARA, uma análise no grande grupo das trinta primeiras ocorrências para o verbo 'get' foi realizada, observando-se quais sentidos eram mais frequentes. Em seguida, os resultados foram comparados com os significados fornecidos pelo Google Tradutor. Concluída a pesquisa, a etapa final do experimento passa a ser realizar novas buscas com outros verbos e, desta vez, em duplas, os alunos pesquisam o sentido de outros verbos com múltiplos sentidos que geralmente são mais complexos de adquirir por eles (os verbos propostos seria o make, do e take). Sugere-se, por fim, que os aprendizes de inglês como língua estrangeira registrem suas conclusões antes de discutirmos sobre os resultados. O experimento confirmou a importância da linguística de corpus para a o ensino e aprendizagem de inglês no contexto brasileiro. O uso de corpus como ferramenta de ensino de língua estrangeira é salutar, pois permite que o próprio aprendiz interaja com os seus pares e construa seu próprio conhecimento, internalizando de um modo muito mais significativo para ele aquilo que causava certa barreira para avançar em direção à língua-alvo. Ficou evidente que, conforme defende Leech (1986), mesmo corpora pequenos que não contêm todas as possibilidades de experiência com estruturas linguísticas são úteis para engajar o aluno a refletir sobre a sua interpretação e a criar modelos sobre estruturas linguísticas com autonomia. Desse modo, o experimento contribui para facilitar a internalização de padrões léxico-gramaticais (cf. Sardinha, 2006) do verbo get através da experiência com ocorrências de get no corpus. Trabalhos como este contribuem para fortalecer os laços entre a LC e o ensino de língua estrangeira, na medida em que refletem sobre práticas pedagógicas para contextos específicos.

Contato: tarsilabattistella@yahoo.com.br

DIMENSIONS OF VARIATION OF BRAZILIAN PORTUGUESE

Tony Berber Sardinha (LAEL/PUC-SP)
Cristina Mayer Acunzo (LAEL/PUC-SP)
Carlos Kauffmann (FSP)

This paper aims at identifying the dimensions of variation of Brazilian Portuguese by means of corpus-based analysis, more specifically by Multidimensional Analysis (MDA; BIBER, 1988 et seq.). MDA makes use of the construct of register, which means 'a cover term for any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances' (BIBER, 2009: 2003), whose scope is flexible and may range from very specific genres such as research articles to broader categorizations such as 'official documents'. Dimensions of variation, in turn, are patterns of cooccurrence of linguistic features underlying the registers of a language (BIBER, 2009). As such, they capture the space of variation across texts and registers, synthesizing it and showing the relative distance among the registers under investigation. An example of dimension of variation (for English) is 'Interaction versus Information' (BIBER, 1988), which indicates that all texts of that language bear these salient characteristics, which are interaction, on the one hand, and information, on the other. Previous research has carried out the identification of dimensions of several languages, such as English (BIBER, 1988; CROSSLEY, LOUWERSE, 2007; DE MÖNNINK, BROM, OOSTDIJK, 2003; LEE, 1999), Korean (KIM, BIBER, 1994), Somali (BIBER, HARED, 1994), Nukulaelae (BESNIER, 1988), Gaelic (LAMB, 2008) and Spanish (BIBER, DAVIES, JONES et al., 2006; PARODI, 2007); however, the multidimensional variation of Portuguese had not been carried out, despite the fact that Portuguese is an important European language, the second largest Romance language. The Brazilian variety accounts for 90% of its native speakers. We intend to fill this gap (for the Brazilian variety of Portuguese) with the current proposal. This project is funded by both Fapesp and CNPq, and its goals are: (1) Find the relevant lexico-gramatical variables that may explain the textual variation of Brazilian Portuguese; (2) Experiment with other types of variable such as collocational, semantic and keywords to verify their suitability for the study of multidimensional variation of Brazilian Portuguese; (3) Identify the statistical factors that underlie textual variation of Brazilian Portuguese as well as the share of variance accounted for by each factor; (4) Interpret the factors by means of criteria that are discursive, functional or communicative in nature; (5) Propose dimensions of variation based on the factors extracted; (6) Develop and make available resources for disseminating and popularizing MDA, enabling other researchers to carry out their own analyses. The methodology is the following: (1) A 2 million word sample of the 1-billion-word Brazilian Corpus was chosen, comprising major spoken and written registers; (2) The corpus was tagged for part of speech using the Palavras Parser; (3) Counts were taken for each feature by scripts specially developed for this purpose by the research team; counts were then normalized, and standardized; (4) An initial Factor Analysis was run on SPSS, and the number of factors in the data were established; (5) A subsequent rotated FA was run for the specified number of factors; (6) Factors scores were computed for each text on each factor; (7) Factors were interpreted in terms of underlying dimensions of variation. Results show a number of different factors that were interpreted, each representing a particular aspect of register variation in Brazilian Portuguese. It is not the intent of MDA to oppose specialty areas such as Genre

Analysis or Discourse Analysis, given that these have their own goals; MDA seeks to provide an empirical large-scale view of central issues surrounding register variation in society.

Contato: tony@corpuslg.org

Referências:

BESNIER, N. The linguistic relationships of spoken and written Nukulaelae registers. Language, v. 64, p. 707-736, 1988.

BIBER, D. Variation across Speech and Writing. Cambridge: Cambridge University Press, 1988.

BIBER, D. Multi-dimensional approaches. In: LÜDELING, A.; KYTÖ, M. (Ed.). Corpus Linguistics -- An International Handbook. Berlin / New York: Walter de Gruyter, 2009.

BIBER, D. et al. Spoken and written register variation in Spanish: A multi-dimensional analysis. Corpora, v. 1, n. 1, p. 1-37, 2006.

BIBER, D.; HARED, M. Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers. In: BIBER, D.; FINEGAN, E. (Ed.). Sociolinguistic Perspectives on Register. Oxford: Oxford University Press, 1994. p. 182-216.

CROSSLEY, S.; LOUWERSE, M. M. Multi-dimensional register classification using bi-grams. International Journal of Corpus Linguistics, v. 12, n. 4, p. 453-478, 2007.

DE MÖNNINK, I. M. et al. Using the MF/MD method for automatic text classification. In: GRANGER, S.; PETCH TYSON, S. (Ed.). Extending the scope of corpus based research: new applications new challenges. Amsterdam: Rodopi, 2003. p. 15-25.

KIM, Y.-J.; BIBER, D. A corpus-based analysis of register variation in Korean. In: BIBER, D.; FINEGAN, E. (Ed.). Sociolinguistic Perspectives on Register. Oxford: Oxford University Press, 1994. p. 157-181.

LAMB, W. Scottish Gaelic speech and writing: register variation in an endangered language. Belfast: Cló Ollscoil na Banríona, 2008.

LEE, D. Y. W. Modelling Variation in Spoken and Written Language: the Multi-Dimensional Approach Revisited. (Tese de doutoramento), Department of Linguistics and Modern English Language, Lancaster University, UK, 1999.

PARODI, G. Variation across registers in Spanish: Exploring the El-Grial PUCV Corpus. In: PARODI, G. (Ed.). Working with Spanish Corpora. London: Continuum, 2007. p. 11-53.

PÔSTERES

ENSINO COM LINHAS DE CONCORDÂNCIA: HIGHLY SUCCESSFUL?

Andrea Geroldo dos Santos (USP)

The aim of this paper is to present part of our Master thesis, in which we compiled a 2,310,143-word monolingual corpus (British and American English) composed by business papers and magazines as well as business reports in order to: (1) research the most frequent adverbial collocations in the area and analyze them; (2) prepare exercises to teach these collocations to ESL students of different levels. Here we show how we prepared the exercises by using not only concordance lines of adverbial collocations (such as work closely), but also exploring such lines in written and oral tasks. The paper concludes with observations about the importance of using concordance lines for ESL teaching in a meaningful way.

Contato: andrea.geroldo@gmail.com

RASTREAMENTO DA PRESENÇA DISCURSIVA DO TRADUTOR NO TEXTO TRADUZIDO

Carolina Barcellos (UFMG)
Célia Magalhães (UFMG)

Os estudos descritivos da tradução, mais especificamente o sub-ramo dos estudos da tradução baseados em corpus, têm incluído a análise do estilo de tradutores profissionais e literários como foco de interesse, abordando a presença discursiva do tradutor no texto traduzido. Esses trabalhos têm ainda analisado diversos corpora a fim de identificar possíveis características dos textos traduzidos. Buscando contribuir para o fortalecimento desses estudos, a presente pesquisa analisa traços estilísticos de tradutores, através do rastreamento de categorias da apresentação da fala, escrita e pensamento, no corpus paralelo pequeno-médio formado pela novela Heart of Darkness, de Joseph Conrad, e duas de suas traduções para o português brasileiro, Coração das Trevas, de Sergio Flaksman, e No Coração das Trevas, de José Roberto O'Shea. A anotação sistemática e detalhada de um corpus para descrever categorias de apresentação da fala, escrita e pensamento fornece informações importantes sobre a apresentação do discurso em narrativas e pode demonstrar como esses padrões variam (SEMINO & SHORT, 2004). As perguntas que motivaram este trabalho foram propostas por Baker (2000) e questionam, especificamente, se o tradutor preferiria determinadas estruturas linguísticas independentemente do estilo do autor; se essas escolhas estariam relacionadas ao sistema linguístico em que ocorrem; e, em caso afirmativo para ambas, se as preferências identificadas no texto traduzido poderiam ser justificadas pelo posicionamento ideológico do tradutor. O arcabouço teórico adotado é o dos estudos da tradução baseados em corpus. A identificação dos padrões de escolha nos textos traduzidos, atribuídos ao estilo dos tradutores e não a imposições linguísticas do par inglês/português, consideram o que foi postulado por Toury (1980, 2000) e Baker (1993, 1996, 2000). As ferramentas de análise linguística apontadas até aqui dialogam ainda com o estudo de narrativas ficcionais traduzidas. Baker (2000)

afirma que a noção de estilo sempre esteve relacionada à concepção de "escritos originais", e portanto era sistematicamente desconsidera pelos estudos da tradução. A tradução vista como uma atividade derivativa e, portanto, desassociada do uso criativo da linguagem, assume a premissa de que o estilo do autor deve ser reproduzido fielmente na língua de chegada. Considerando que o ato tradutório envolve a interpretação e subsequente transposição de significados entre dois sistemas linguísticos distintos, é possível questionar a existência de uma total imparcialidade por parte do tradutor. Este trabalho conta ainda com subsídios do modelo desenvolvido por Semino & Short (2004) para rastrear a apresentação de atos de fala, escrita e pensamento e subcategorias em textos originais em inglês. Os procedimentos metodológicos adotados incluem a caracterização do corpus em dados quantitativos através do software Wordsmith Tools© 5.0 e a anotação manual das categorias de classificação da apresentação da fala, escrita e pensamento e subcategorias a partir do modelo descrito por Semino & Short (2004). Os resultados apontam que a fala foi o tipo de discurso mais frequente no corpus e indicam que a análise das orações introdutórias de elocução, em particular, pode revelar traços de estilo em narrativas ficcionais traduzidas para a língua portuguesa. As ocorrências das categorias de apresentação da fala, escrita e pensamento e subcategorias identificadas nos textos traduzidos, em comparação com aquelas encontradas no texto original, apresentam maior variação, em números absolutos, entre ocorrências de categorias de um mesmo tipo de discurso que entre ocorrências de tipos de discurso diferentes. Foram verificadas escolhas distintas entre os tradutores os quais, através de omissão ou explicitação principalmente, introduziram mudanças no texto traduzido em relação ao texto original, sendo, no entanto, prematuro atribuí-las a um posicionamento ideológico por parte dos tradutores. Foi constatado ainda que uma das características do texto traduzido, nomeadamente a explicitação, pode estar estreitamente relacionada ao estilo do tradutor variando sua ocorrência entre os textos traduzidos do corpus analisado. Os resultados obtidos apontam, em geral, que o estilo do autor parece exercer influência nas escolhas feitas pelos tradutores, não sendo, no entanto, determinante.

Contato: cpbarcellos@gmail.com

CORPUS FALADO X CORPUS ESCRITO: DIFERENÇAS QUANTITATIVAS E/OU QUALITATIVAS?

César Nardelli Cambraia (UFMG)

A lingüística histórica conta essencialmente com textos escritos como corpora para a investigação da mudança lingüística. Naturalmente a principal questão é em que medida os textos escritos refletem a língua falada de uma dada época. Diferentes estudos (Teixeira, 2003; Berlink, 2007) têm analisado textos de teatro (sobretudo comédias) em língua portuguesa, concebidos para serem representados oralmente e em função disso, em tese, mais transparentes em relação à língua falada, para responder a essa questão: os resultados tem mostrado que, diferentemente do esperado, o referido gênero não é tão transparente como se supunha. No presente trabalho, investiga-se o comportamento lingüístico dos demonstrativos em corpora em língua portuguesa (variedade brasileira [=PB]) e em língua espanhola (variedade mexicana [= EM]), confrontando-se dados de língua oral com dados de língua escrita, a fim de avaliar de que forma testemunham o processo de simplificação do sistema em que a forma de 1ª pessoa (este) se perde no português

do Brasil (Câmara Jr, 1971) e a forma de 3ª pessoa (aquel) se perde no espanhol do México (Kany, 1994). O corpus de língua falada deriva do Projeto NURC do Rio de Janeiro (Callou & Lopes, 1993), coletado entre 1971 e 1974, e da Cidade do México (Lopez Blanch, 1971), coletado por volta de 1968, tendo sido considerado, em ambos os casos, os três tipos de inquérito (diálogo entre informante e documentador [=DID], diálogo entre informantes [=D2] e elocuções formais [=EF]); o corpus de língua escrita compõe-se de peças de teatro, de autor da mesma localidade do corpus oral: Um elefante no caos: uma farsa [=ELE], de Millôr Fernandes, composta em 1955; e Debiera haber obispas [=DEB], de Rafael Solanas, composta em 1956). Foram coletadas e analisadas as 150 primeiras ocorrências de demonstrativos (1ª pessoa [=F1], 2ª pessoa [=F2], 3ª pessoa [=F3]). Os dados apurados apresentam os seguintes valores: PB = NURC/RJ – F1 5,2%, F2 72,3% e F3 22,5% × ELE – F1 23,5%, F2 70,6% e F3 5,9%; EM = NURC/MX - F1 37,3%, F2 58,3% e F3 4,5% × DEB - F1 45,3%, F2 49,4% e F3 5,3%. Esses primeiros números indicam que a relação entre língua falada e língua escrita difere quanto à variedade lingüística: no PB, a tendência a desaparecimento de F1 é nítida na língua falada (NURC 5,2%) mas não se manifesta claramente na língua escrita (ELE 23,5%); já no EM, a tendência a desaparecimento de F3 é nítida na língua falada (NURC 4,5%) e na língua escrita (ELE 5,3%). Poder-se-ia argumentar que, dos três tipos de inquérito, seria a D2 que o texto teatral mais deveria se assemelhar, pois ambos se caracterizam por diálogos (entre informantes e entre personagens). Considerando para o NURC apenas os dados de D2, tem-se: NURC/RJ – F1 1,3%, F2 61,7% e F3 37%; NURC/MX – F1 21,6%, F2 66,4% e F3 12,1%. Levando em conta apenas o tipo de inquérito mencionado, a discrepância em F1 entre língua falada (1,3%) e escrita (23,5%) aumenta no PB; e a em F3 passa a existir no EM (língua falada, 12,1%; língua escrita, 5,3%). No caso do PB, ocorre algo que não surpreende: a língua escrita mostra-se mais conservadora do que falada; mas no caso do EM a questão se complica, pois a língua falada se mostra mais conservadora do que a escrita! No caso do EM, a discrepância pode ser explicada com base nos informantes: nos dois inquéritos analisados de D2, em três informantes F3 quase inexiste (D2013, informantes A e B, nenhuma ocorrência; D2016, informante B, 1 ocorrência), ficando concentrada em apenas um informante (D2016, informante A, 13 ocorrências), cuja formação é na área de Direito (em que, como se sabe, predomina uma linguagem bem conservadora). Colocando à parte os dados do informante discrepante, o resultado para o EM passa a ser o esperado: a língua escrita volta a apresentar maior porcentagem de F3 do que a falada – trata-se, portanto, de uma diferença quantitativa. Entretanto, verifica-se no corpus que há diferenças também qualitativas: destaca-se, no confronto entre os dados de língua falada e língua escrita no PB, o fato de constatar-se posposição de demonstrativos nos dados de língua falada (0,5%), mas nenhuma no de língua escrita; já no corpus do EM, sobressaem os fatos de que (a) nos dados tanto de língua falada (0,5%) quanto de língua escrita (0,7%) haja a posposição de demonstrativo com anteposição de artigo ao nome e (b) haja uso fático apenas nos dados de língua falada – trata-se, portanto, de diferenças qualitativas. De forma geral, os dados analisados indicam que: (i) há diferenças quantitativas entre o corpus de língua falada e o de língua escrita, o que significa que, mesmo peças teatrais cômicas, apresentam certo grau de conservadorismo e, por isso, a freqüência dos padrões lingüísticos constatada nestas últimas deve ser relativizada, sobretudo em uma análise diacrônica em que se correlaciona um dado fenômeno com outros fenômenos intralingüísticos (como a adjunção de advérbios como estratégia para maior definição do valor dos demonstrativos) e extralingüísticos (como mudanças sociais); (ii) há também diferenças no grau de distância entre língua falada e escrita em distintos domínios lingüísticos (a língua escrita do PB mostrou-se mais conservadora do que a do EM), o que significa que a relativização das freqüências

não pode ser calculada com base em uma fórmula única para qualquer domínio lingüístico; e (iii) as diferenças não se dão apenas em termos quantitativos, mas também qualitativos, pois o corpus oral apresenta padrões estruturais distintos do escrito (atente-se para a posposição de demonstrativos apenas no corpus oral do PB e para o uso fático apenas no corpus oral do EM). Em síntese, a relação entre língua oral e língua escrita precisa ser investigada mais a fundo para se identificarem as peculiaridades segundo a natureza do corpus (sejam as diferentes formas de registro de língua oral, como diálogo entre informante e documentador, diálogo entre informantes, etc., sejam as diferentes formas de registro da língua escrita, como peças de teatro, romances, narrativas, etc.) e também segundo o domínio lingüístico (do português, do espanhol, etc.).

Contato: nardelli@ufmg.br

UM ESTUDO DE CORPUS NAS METÁFORAS DO CONCEITO SOCIEDADE EM ALEMÃO

Emanuela Costa (UFMG)

Contextualização: Desde o lançamento do livro Metaphor we live by (LAKOFF; JOHNSON, 1980), a metáfora deixa de ser vista apenas como uma figura da retórica, mas como parte do nosso sistema conceitual, ou seja, o sistema orientador do nosso pensamento. Desse modo, o uso das metáforas está relacionado ao modo como entendemos vários conceitos existentes na língua, pois a operação cognitiva que se processa é o emprego do domínio-fonte, mais experiencial, com o domínio-alvo, mais abstrato, assim entendemos o conceito mais abstrato, por meio de um mais concreto (i.e., B por meio de A), o que demanda o uso de uma metáfora (LIMA, 2001). Logo, para se falar de um termo puramente abstrato com sociedade, a mente faz elaborações complexas, que podem ser mapeadas a luz da teoria cognitiva da metáfora. De acordo com o estudo feito em Schröder 2009, sobre a construção metafórica da sociedade, concluiu-se que havia diferença na conceitualização do conceito sociedade em português e alemão. O estudo revelou uma tendência do corpus em alemão em metaforizar esse conceito por meio de esquemas imagéticos misturados e dinamizados, além das metáforas conceituais encerrarem em muitos casos, como domínio fonte os conceitos de negócios, prédios, jogos e observação. O corpus em português, por outro lado, tendia a ser motivado pelos domínios-fonte da família, da guerra, flora e estágio, além de ter sido mais personificado. Contudo, o estudo apresentado pretende analisar como o esquema da verticalidade, do centro-periferia, e o esquema do caminho na construção do conceito sociedade são metaforizados, em textos jornalísticos em português e em alemão. A análise dos dados será feita por meio de uma preparação de 200 textos em alemão, retirados da revista "Der Spiegel" e 200 textos em português, retirados da revista "Carta Capital". Considerando essas metáforas conceituais, os resultados explicarão alguns dos padrões dessas línguas, pois o exame detalhado dos resultados poderá implicar no conhecimento de como o pensamento dessas duas culturas se organiza. A pesquisa buscará descrever e comparar um grupo de sentenças relacionadas às metáforas conceituais descritas acima. Como se trata de um estudo intercultural, poderemos avaliar como as metáforas conceituais semelhantes são projetadas, de modo a representar as diferenças culturais. Revisão de Literatura: Uma das mais fortes afirmações da teoria cognitiva da metáfora é a importância que as experiências físicas fundamentam as muitas das metáforas conceptuais mais básicas. No entanto, eventos sociais que ocorrem nas culturas estudadas

também podem produzir metáforas, já que textos jornalísticos podem conter comentários feitos por políticos e jornalistas que servem de input para a produção metafórica conceptual, já que criam uma certa intertextualidade em seus domínios sociais, comprovando assim o caráter criativo e inovador da metáfora (Schröder 2009). Objetivo do trabalho: Projeta-se que, com esta pesquisa, será possível entender melhor o processo cognitivo que leva às diferentes organizações dos domínios que servem como inputs para a produção metafórica conceitual e como eventos históricos influenciam nessa produção. Portanto, este projeto visa, dentre outros aspectos, explorar mais profundamente a relação da cultura dentro desse processo cognitivo que gera as metáforas. Mais especificamente, isso será feito a partir da verificação, de como o termo sociedade é metaforizados em português e alemão, em um corpus midiático; Der Spiegel, Carta Capital. A pesquisa buscará descrever e comparar um grupo de sentenças relacionadas às metáforas conceituais descritas acima. Como se trata de um estudo intercultural, poderemos avaliar como as metáforas conceituais semelhantes são projetadas, de modo a representar as diferenças culturais. Desse modo, esse estudo intercultural deseja investigar como funciona a organização dos colocados, que representam os esquemas imagéticos verticalidade, centro-periferia, caminho nos corpora. Metodologia: Preparação de um corpus em português e outro e alemão. O corpus em português será formado de 200 textos retirados do site da revista Carta Capital e depois de 200 textos retirados do site da revista Der Spiegel. Os dados serão analisado no programa AntCon e em seguida os dados serão apresentados. Conclusão: A dificuldade que podemos enfrentar será a provável falta de robustez dos dados, já que apesar do corpus se compor de dois jornais, em alemão e outro em português, não podemos afirmar se a diferença de tamanho dos corpora influencia no resultado.

Contato: emanuela.costa@gmail.com

A Construção Superlativa De Expressão Corporal: Uma Análise Baseada Em Corpora

Igor Costa (UFJF)
Neusa Salim Miranda (UFJF)

Este trabalho tem como objeto a Construção Superlativa de Expressão Corporal (Construção SEC) (“solteirona e toda virgem, ignorava machezas, quase morreu de vergonha numa tarde de conversas”; “Padre Dito quase estourou de rir”; “O Lúcio rolou de rir com a explicação, e como consequência acabou virando a vítima e a cobaia do seminário”), aqui postulada como um nódulo da grande rede de construções do Português denominadas por Miranda (2008) Construções Superlativas. O enfoque teórico advém da Linguística Cognitiva (FAUCONNIER, 1997; FAUCONNIER E TURNER, 2002; FILLMORE, 1977, 1982; FILLMORE E ATKINS, 1992; JOHNSON, 1987; LAKOFF, 1987, 1993; LAKOFF E JOHNSON, 1980, 1999; MIRANDA, 2002, 2008; SALOMÃO, 2002, 2008; dentre outros) que entende a linguagem como uma faculdade cognitiva não-autônoma regulada por aparato cognitivo geral; advoga um papel central para processos imaginativos (metáfora, metonímia, mesclagem) na cognição e na linguagem humanas; vê a gramática como conceptualização; e entende que o conhecimento sobre a linguagem emerge do uso. Mais especificamente, serve de endosso a este estudo um modelo de gramática desenvolvido no interior do paradigma cognitivista, a Gramática das Construções Cognitiva (GOLDBERG, 1995,

2006; CROFT e CRUSE, 2004; GOLDBERG, 1995, 2006; LAKOFF, 1987; TOMASELLO, 2003[1999]; BOAS, no prelo) que, definindo construções como pares de forma-função, confere-lhes o estatuto de unidades básicas da linguagem. Nesse enquadre, a gramática e o léxico se definem como uma rede de construções instituídas pelo uso através da cultura. Partindo desses pressupostos, propõe-se a descrição dos padrões formais e semântico-pragmático emergentes em ocorrências dessa construção, de modo a se elucidarem a configuração do par forma-sentido e os processos cognitivos definidores do padrão abstrato sob estudo. Dado o relevo do uso no modelo teórico-analítico adotado, opta-se metodologicamente por uma análise com base em corpora (ALUÍSIO E ALMEIDA, 2006; GRIES E DIVJAK, 2003; SARDINHA, 2000, 2004; STEFANOWITSCH, 2006), devido à possibilidade de se lidar com a linguagem situada no discurso real e quantificá-la. A Gramática das Construções Cognitiva, como um Modelo Baseado no Uso (GOLDBERG, 1995, 2006; CROFT e CRUSE, 2004), vê a linguagem de um viés empirista, entendendo-a como um sistema probabilístico, fazendo com que a frequência de ocorrência (token) e de tipos (types) tenham papel central na definição da convencionalização e produtividade da construção. O corpus utilizado para pesquisa é o Corpus do Português (http://www.corpusdoportugues.org/), composto por quarenta e cinco milhões de palavras, distribuídas em textos que perpassam os séculos XIV-XX. Partindo do trabalho de Sampaio (2007), em que parte da Construção SEC foi analisada dentro dos limites do campo metafórico da "morte" ("morrer de rir", "morrer de medo"), nossas análises ampliam tal estudo, investigando a produtividade e a convencionalização dessa rede e a natureza do desencontro ou discrepância semântico e sintático (o fenômeno do mismatch/desencontro) que institui, sincronicamente, seu padrão. Tal padrão – [XV de YN/V] – tem em X um verbo ("chorar de", "rolar de", "morrer de", "se acabar de", "se arrebentar de", dentre outros) que representa o impacto físico ou fisiológico metafórico desencadeado pelo excesso de Y, um SN de natureza abstrata ("de medo", "de tristeza") ou um SV ("de rir", "de estudar"). Os principais pontos em nosso quadro descritivo são os seguintes: (i) desencontro na semântica de XV (verbos que suscitam os frames de impacto físico ou fisiológico), que, na construção tem função de Operador Escalar (morrer de medo, rolar de rir); (ii) condição de semiauxiliar modal de XV, quando Y é um verbo; (iii) uso pragmático da construção como estratégia argumentativa pertinente a contextos discursivos em que o falante/escritor possui maior liberdade de expressão subjetiva; (iv) centralidade de processos figurativos (especialmente metafóricos) na instituição do padrão construcional. As metáforas primárias "Causa É Força Física" e "Intensidade É Escala" (LAKOFF E JOHNSON, 1999) atuam como bases conceptuais do padrão construcional. Da análise da freqüência dos dados emerge a afirmação da produtividade da Construção SEC, que instancia, nos corpora investigados, 19 diferentes types. O processo de convencionalização se delineia pela presença de 1.726 tokens, sendo "Cansar(-se) de Y", "Cair de Y", "Chorar de Y", "Fartar(-se) de Y" e "Morrer de Y" os types mais convencionalizados. Tais achados analíticos, desvelando as especificidades deste padrão, legitimam a postulação da Construção SEC como uma construção do Português, com forma, sentido e uso próprios, e como um elo da grande rede de Construções Superlativas dessa língua. Considere-se como ganho substancial deste estudo a abordagem metodológica eleita que, ancorada em corpora, trouxe para nossas análises informações que emergiram naturalmente dos dados, permitindo, assim, descrições mais precisas do mesmo que ultrapassaram possíveis respostas cunhadas apenas pelo nosso julgamento intuitivo de pesquisador. Ganha a Linguística Cognitiva que, tendo o uso como condição analítica de suas unidades construcionais, opta, presentemente, por uma Linguística baseada em corpus.

Contato: igorsabo@yahoo.com.br

CHICK LIT, CHICK MAGNET, BIKER CHICK, HIPPIE CHICK: METAPHORICAL USES OF CHICK IN CONTEMPORARY AMERICAN ENGLISH

Lívia Pretto Mottin (PUC-RS)

Metaphors are all around us: newspapers, magazines, television, everyday conversations, internet, books, and so on. One of the most important achievements in metaphor theory in the last years is that it is not just a poetic or rhetorical tool, but it is part of our conceptual system, of the way we think and act (Lakoff and Johnson, 1980). It allows us to understand one aspect of a concept in terms of another. However, metaphors can sometimes be confusing and not so easily understandable for learners of a foreign language. As the interpretation of the metaphorical expressions depends on the active participation of the receiver, he/she has to be actively involved in order to be able to realize the meaning behind the expression. Nevertheless, sometimes, even though the receiver is attentive, establishing the appropriate meaning and sense of metaphorical expressions may be hard. This paper aims at unraveling how the lexeme chick can be used to refer to women and to the female universe. I will first briefly review some literature on metaphor and, then, check how the domain of animals can be used to better explain human behavior and how the use of electronic corpora and software tools changed the context of metaphor studies providing other manners of analysing metaphors. Then, I will present some data collected in the Corpus of Contemporary American English, a 425 million words corpus, for identifying and interpreting some metaphors of the lexeme chick through a corpus-based analysis. After defining the search word, chosen for its high metaphorical potential, I first searched for the main collocate terms (one position to the left and one position to the right) of the lexeme chick, in order to observe if the meanings were figurative or not. Next, from the total number of occurrences of the search word, I checked the distribution of chick among the five subcorpora, aiming at investigating in which language variety of COCA the lexeme appears more frequently. Subsequently, by using the KWIC (Key Word in Context) concordance format, the first 100 random concordance lines in each subcorpus were read and carefully examined, in order to determine whether the lexeme was used in its literal or metaphorical sense. The non-literal occurrences were observed in more detail aiming at finding out the underlying metaphorical use they represent. Afterwards, I investigated all the figurative occurrences, so that I could determine the positions the lexeme may occupy in a sentence, i.e. if chick was used as a noun or as an adjective, and to verify its strongest collocations. With regard to the search word chick in its metaphorical sense, it is possible to conclude that: (i) it is used to refer to women and to other things that make part of the female universe; (ii) the target domain WOMEN is understood in terms of the source domain CHICK; (iii) it is very frequent in informal situations, especially in spoken, fiction and popular magazines subcorpora; (iv) it is not very frequent in the newspaper subcorpus; (v) it is rare in academic language; (vi) it is used either as a noun or as an adjective; (vii) as a noun, it produces the collocations biker chick and hippie chick; (viii) as an adjective, it produces the collocations chick flick, chick lit, and chick magnet; (ix) the Oxford Advanced Learners Dictionary (2005) brings just two of the collocations revealed by COCA (chick flick and chick lit); (x) it is a productive metaphor; (xi) the occurrences of chick in its metaphorical sense outnumber the occurrences in its literal sense. Without electronic corpora, it

would be very difficult to get to these conclusions. Although Corpus Linguistics has developed a lot in the last years, when concerning metaphors researches, there are still some data which are not reachable through corpora and software. Chick, the search word of my research, is an example of how potentially metaphorical a lexeme can be and the data collected in COCA confirm its omnipresence in our day by day and the importance of teaching metaphors in a foreign language learning context, since they make part of real English.

Contato: liviamottin@yahoo.com.br

Referências:

BERBER SARDINHA, A. P. Metaphor in corpora: A corpus-driven analysis of Applied Linguistics dissertations. Revista Brasileira de Linguística Aplicada. Volume 7, n. 1, 2007, p. 11-35.

BERBER SARDINHA, A. P. Análise de metáfora em corpora. Ilha do Desterro, A Journal of English Language, Literatures in English and Cultural Studies, No 52, 2007, p. 167-199.

DAVIES, M. The Contemporary Corpus of American English. Available at: http://corpus.byu.edu/coca/

HORNBY, A. S. Oxford Advanced Learners Dictionary. Oxford: Oxford University Press, 2005.

KÖVECSES, Z. Metaphor: A Practical Introduction. New York: Oxford University Press, 2010.

LAKOFF, G.; JOHNSON, M. Metaphors we live by. Chicago: The University of Chicago Press, 1980.

LAKOFF, G. The contemporary theory of metaphor. In: ORTONY, A. (Ed.), Metaphor and thought. Second edition. Cambridge: Cambridge University Press, 1993, p. 202-251.

LAKOFF, G. Women, Fire, and Dangerous Things – What Categories Reveal about the Mind. Chicago: The University of Chicago Press, 1987.

McENERY, T; WILSON, A. Corpus Linguistics: an Introduction. Edinburgh: Edinburgh University Press, 2004.

RODRÍGUEZ, I. L. Of Women, Bitches, Chickens and Vixens: Animal Metaphors for Women in English and Spanish. Revista de Estudios Culturales de la Universitat Jaume I / Cultural Studies Journal of Universitat Jaume I, vol. 2 , 2009, p. 77-100.

RODRÍGUEZ, I. L. The representation of women in teenage and women's magazines: recurring metaphors in English. Estudios Ingleses de la Universidad Complutense, vol. 15, 2007, p. 15-42.

ARQUITETURA E CRIAÇÃO DE UM CORPUS PARA O ESTUDO DA EROSÃO LINGUÍSTICA

Lucia A. Ferrari (PosLin/UFMG)

O trabalho mostra a aplicação da linguística de corpus (LC) ao estudo da erosão da L1, e discute questões metodológicas delicadas a respeito. Entende-se por erosão linguística a re-estruturação gradual, convergência ou perda de estruturas fonológicas, morfossintáticas, lexicais e pragmáticas na produção de falantes de L1, em contato prolongado com uma L2, devido à interferência desta,

ou por falta de insumo (Schmidt et alii, 2004). Neste trabalho a L1 estudada é o italiano, enquanto a L2 é o português brasileiro (PB). Os trabalhos sobre erosão linguística que se utilizam da metodologia de corpus são escassos: geralmente são empregados testes de tradução, julgamento sobre a gramaticalidade de enunciados, questionários sociolinguísticos, C-Tests com completamento de lacunas, narrativas dirigidas ou narração de trechos de um filme mudo, Wug Tests e Can-do Scales. Acreditamos que, para obter informações o mais espontâneas possíveis, é imprescindível o uso da LC, que permite por um lado o acesso à língua realmente usada pelos informantes e por outro a comparação com o uso dos informantes não sujeitos a erosão. Os estudos mais influentes sobre erosão linguística concentram-se em Seliger e Vago (1991) e Köpke e Schmid (2004), além de trabalhos publicados nas revistas International Journal of Bilingualism (em específico nos referimos aqui ao Vol 8:3, 2004) e em várias edições da revista Bilingualism: Language and Cognition. Além destes, é fundamental Keijzer (2007). Contudo, raramente é utilizada a metodologia da LC: Keijzer a utiliza somente sob forma de narrativas guiadas, entre muitos outros testes que considera mais relevantes. Hutz (2004) faz um estudo longitudinal sobre um corpus de cartas, mas trata-se de um corpus escrito. Os estudos, Raso e Vale (2007 e 2009) concentram-se sobre um corpus de 18080 palavras de língua falada: trata-se de uma série de pesquisas sobre a erosão linguística de italianos cultos em contato prolongado com o PB nas quais a metodologia da LC é parte essencial. O único outro trabalho de nosso conhecimento que se utilize desta metodologia e que aborde o contato de uma língua estrangeira com o PB é Calvo Capilla (2007). O objetivo de nosso estudo foi verificar os resultados das pesquisas de Raso e Vale utilizando um novo corpus, de 21298 palavras, que seguisse critérios mais rigorosos e que fosse comparado com textos semelhantes, para um total de 21224 palavras, de um corpus de monolíngues alinhado (permitindo assim o acesso imediato ao som) e criado com uma arquitetura e uma variação diafásica sofisticadas, o C-ORAL-ROM (Cresti e Moneglia 2005). Foram analisados os mesmos itens pesquisados por Raso e Vale: os pronomes clíticos ci em seus valores atualizante, lexicalizante e locativo, os clíticos ne em suas funções partitiva, argumental e locativa e os clíticos acusativos de terceira pessoa lo, la li, le, l'. Para a criação do novo corpus e as escolhas sobre sua arquitetura foram seguidas as indicações de Biber (1993), McEnery; Wilson (1996) e Meyer (2004), entre outros. Os passos seguidos para a compilação foram os seguintes: Escolha dos informantes adequados: italianos nativos crescidos na Itália até a idade adulta e que ali tivessem completado pelo menos até o segundo grau, que é de cinco anos, e possivelmente possuíssem formação universitária. Isto para evitar a ambiguidade entre erosão e aquisição incompleta e para contar com informantes que tivessem uma adequada capacidade de reflexão metalinguística. Estes deveriam residir no Brasil há pelo menos oito anos, período tradicionalmente considerado suficiente para a detecção dos fortes sinais da erosão linguística. Gravação com equipamentos sem fio adequados para garantir uma qualidade acústica suficiente para a análise prosódica. Escolha de interações as mais variadas possíveis, evitando aquela da entrevista, diafasicamente pouco significativa. Transcrição em formato CHAT das gravações e seleção dos trechos a serem utilizados. Varredura do corpus de forma manual à procura dos itens objeto de estudo. Análise dos resultados. Os resultados confirmaram em boa medida o que foi encontrado nos estudos anteriores, mas o que conta mais é que as divergências mostraram-se extremamente úteis para a reflexão metodológica. O corpus analisado por Raso e Vale é composto principalmente por entrevistas com um assunto definido; o novo corpus privilegia interações diafasicamente variadas. Isto favoreceu a produção de estruturas diferentes nos dois corpora. Damos aqui um único exemplo, entre os mais importantes no estudo da erosão pronominal e da estruturação do

enunciado: os pronomes anafóricos de constituintes deslocados à esquerda. Esses pronomes mostraram um aumento no corpus analisado por Raso e Vale em 26,02%, enquanto diminuíram em 21,01% naquele atual. De fato a função pragmática dos deslocamentos à esquerda é uma função temática (ou de topic) que identifica o âmbito de aplicação da força ilocucionária em comentário. Em interações de tipo tendencialmente monológicas, pouco ligadas ao contexto situacional, essa função se torna fundamental para comunicar ao interlocutor o âmbito semântico ao qual o ato linguístico se refere. Ao contrário, em interações de tipo realmente interativo e principalmente onde o contexto situacional joga um papel fundamental, normalmente o âmbito semântico de referência do ato é situacionalmente dado Divergências foram encontradas entre os próprios corpora de comparação: o BADIP (De Mauro, 1993), utilizado por Raso e Vale, composto por interações mais planejadas e o C-ORAL-ROM italiano, muito espontâneo. Entre as discrepâncias citamos o uso do ci locativo. As ocorrências normalizadas a cada 10000 palavras do BADIP somam 21,27, enquanto caem para 8,95 no C-ORAL-ROM italiano. Emergiu também o enorme número de ocorrências, 129,09 a cada 10000 palavras, detectado no C-ORAL-ROM italiano do clítico ci atualizante, em união com os verbos essere e avere, enquanto no BADIP este número é somente de 59,18. O uso do ci atualizante é índice de informalidade e espontaneidade. Em específico o ci com o verbo avere (averci) era considerado não padrão até poucas décadas atrás e evitado também na língua falada. Esses e muitos outros aspectos do trabalho mostram que a LC pode ajudar muito na compreensão dos fenômenos de erosão linguística, mas tornam decisiva uma reflexão atenta sobre a constituição dos corpora de estudo e daqueles de comparação.

Contato: ferrari.lu@gmail.com

Referências:

BIBER, D. Representativeness in Corpus Design. In: Literary and Linguistic Computing, Vol. 8, n. 4, Oxford University Press, 1993.

CALVO CAPILLA, M.C. Espanhol e português em contato: o atrito da L1 de imigrantes espanhois no Brasil. Brasília: 2007. 173 f. Dissertação de Mestrado. Universidade de Brasília, 2007.

CRESTI, E.; MONEGLIA, M. (a cura di) C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Language. Amsterdam-Philadelphia: John Benjamin, 2005.

DE MAURO, T. et alii . Lessico di frequenza dell'italiano parlato. Milano: EtasLibri, 1993.

HUTZ, M. Is there a natural process of decay? A longitudinal study of language attrition. In: KOPKE, B.; SCHMID, M.S. First Language Attrition. Interdisciplinary perspectives on methodological issues. Amsterdam/Philadelphia: John Benjamin, 2004, pp. 189-206.

KEIJZER, M. Last in first out? An investigation of the regression hypothesis in Dutch emigrants in Anglophone Canada. Vrije Universiteit, 2007.

KÖPKE, B.; SCHMID, M.S. First Language Attrition. Interdisciplinary perspectives on methodological issues. Amsterdam/Philadelphia: John Benjamin Publishing Company, 2004.

MACWHINNEY, B. The CHILDES project: tools for analysing talk. Hillsdale: Lawrence Erlbaum, 1994.

MACWHINNEY, B. The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription format and programs. Volume 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

McENERY, T.; WILSON, A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1996.

MEYER C.F. English Corpus Linguistics. An introduction. Cambridge: Cambridge University Press, 2004.

RASO, T. Erosione dei clitici e strutture tematizzanti in italiani colti in contatto prolungato col portoghese brasiliano. In: Sintassi storica e sincronica dell´italiano. Subordinazione, coordinazione, giustapposizione. Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana, 2009. p. 384-399.

RASO, T.; VALE, H. P. A erosão lingüística em italianos cultos em contato prolongado com o português do Brasil: os clíticos e alguns efeitos na estrutura do enunciado. In: Revista de Italianística, v. 16, 2009, p. 1-22.

SCHMID M.S. "A new blueprint for language attrition research". In KÖPKE B. e SCHMID M.S., First Language Attrition. Interdisciplinary perspectives on methodological issues. Amsterdam/Philadelphia: John Benjamin Publishing Company, 2004.

SELIGER H.W. & VAGO R.M. (Eds.), First Language Attrition (pp. 175-188). Cambridge: CUP, 1991.

VALE, H. P. A erosão lingüística dos italianos cultos em contato prolongado com o português do Brasil: os clíticos. Belo Horizonte: Monografia apresentada no Curso de Graduação em Letras da Faculdade de Letras da Universidade Federal de Minas, 2007.

THE JLPT CORPUS: A GUIDANCE TOOL TO JAPANESE STUDENTS

Marco Fonseca (UFMG)

This work aims to compile a corpus consisting of the Japanese Language Proficiency Test (Nihongo Nouryoku Shiken, hence JLPT) previous examinations in order to check the token frequency of the grammar patterns on them. First, we will define what the JLPT is and why to compile its previous tests is important. Then we will explain how we are going to put the corpus together. After that we will show the results we have found so far. Finally, we will discuss future work. The JLPT "is a standardized criterion-referenced test to evaluate and certify the Japanese language proficiency of non-native speakers" (http://en.wikipedia.org/wiki/Japanese_Language_Proficiency_Test). It is held by the Japan Foundation and the Japanese government. The test takes place twice a year in Japan and once a year around the world. It is divided into 5 levels (from N1 to N5 – N stands for Nihongo, Japanese language), each one representing one level of proficiency. N5 corresponds to elementary level and N1, to the advanced. The other tests correspond to the increasing levels of proficiency between these two. The test suffered some changes in 2011 (see http://www.jlpt.jp/e/), but it is basically divided into the following categories: moji (characters), goi (vocabulary) bunpou (grammar) choukai (listening). In this work, we will focus on the grammar section. This section presents small sentences in which the student is required to choose the correct grammar pattern. Sometimes it is not clear for people who are taking this test on which grammar pattern they should focus since the scope of the test is enormous. Preparatory books present about 150 patterns (see Sasaki and Matsumoto, 2010). The search of these patterns in the previous tests is going to help learners to see which patterns are the most frequent. Therefore, it will be possible to see if they are equally distributed among the tests or if some

patterns are more likely to appear in the tests throughout the years. The corpus contains the tests from 1991 to 2009. This is due to methodological and availability reasons. The tests can be freely downloaded from the web in PDF format (see http://nihonhacks.com/japanese-language/past-jlpt-tests/). We converted to txt all the PDFs and then compile the examinations using the Sketch Engine (www.sketchengine.co.uk). This interface contains a segmentation tool for Japanese language, the Chasen. After that, we looked for each one of the 139 grammar patterns presented by Ueki, Ueda and Noguchi (2005) in their preparatory book. The results show that the patterns that appear in the JLPT are not equally distributed in the tests. For example, the pattern katawara (as a sideline, besides) appeared only twice in the tests from 1999 to 2009; however, the pattern koto da shi (since) presented 137 tokens. Therefore, our work can be used as a guidance tool to those who wish to take the JLPT N1 in the future. It can help students to know on what they should focus in order to study for the test. The student will be able to know which patterns they should learn and which ones they should ignore. Future work might consist in a linguistic analysis which can coherently explain the grammar patterns that appears in the test. In other words, it would be interesting to verify if there is, for example, a semantic tendency in the patterns that appeared the most in the previous examinations. It would also be interesting to compile corpus from the other levels of the test (from N1 to N5) in order to see the grammar patterns which appear in the examination as a whole. Furthermore, more work can be done in the other sections of the JLPT (characters, reading, listening).

Contato: marcosilvafonseca@gmail.com

Referências:

JAPANESE Language Proficiency Test. Online. Available at: <http://en.wikipedia.org/wiki/Japanese_Language_Proficiency_Test>. Access on 06/18/2011.

JAPANESE Language Proficiency Test. Online. Available at <http://www.jlpt.jp/e/> . Access on 06/18/2011.

JAPANESE Web Corpus. Online. Available at: <http://the.sketchengine.co.uk/>. Access on 06/18/2011.

PAST JLPT tests. Online. Available at <http://nihonhacks.com/japanese-language/past-jlpt-tests/> . Access on 06/18/2011.

SASAKI, Hitoko. MATSUMOTO, Noriko. Nihongo nouryoku shiken – Nihongo sou matome N1 bunpou (Japanese Language Proficiency Test - Overall of compilation of the N1 grammar). ASK: Tokyo, 2010.

TOMOMATSU, Etsuko. MIYAMOTO, Jun. Wakuri, Masako. Donna toki dou tsukau nihongo (How and when to use Japanese language). ALC: Tokyo, 2007.

UEKI, Kaori. UEDA, Sachiko. NOGUCHI, Kazumi. Kanzen master: 1 kyuu nihongo nouryoku shiken bupou taisaku (Master it completely: a preparatory book for the Japanese Language Proficiency Test level 1). 3A Network: Tokyo, 2005.

UMA FERRAMENTA DE APOIO AOS ESTUDOS DA SONORIDADE SEGMENTAL

Thaïs Cristófaro Silva (UFMG)
Gustavo Mendonça (UFMG)
Leonardo Almeida (UFMG)
Eduardo Gonçalves (UFMG)

A Teoria de Exemplares (Johnson, 1997; Pierrehumbert, 2001; Foulkes e Docherty, 2006) e a Fonologia de Uso (Bybee, 2001) fazem predições específicas quanto a organização segmental da cadeia sonora da fala, sobretudo, quanto a fenômenos de variação e mudança sonora. Estas teorias sugerem que mudanças sonoras que tenham motivação fonética afetem inicialmente os itens lexicais mais frequentes da língua. Por outro lado, mudanças sem motivação fonética atingiriam inicialmente as palavras menos freqüentes da língua. Se efeitos de freqüência são relevantes para a testagem de teorias que investigam aspectos segmentais da linguagem faz-se importante construir uma ferramenta que implemente buscas de padrões segmentais específicos. Este trabalho apresenta uma ferramenta de buscas de padrões segmentais que tem por objetivo oferecer para a comunidade científica um mecanismo de buscas para testar os efeitos de freqüência sugeridos pela Teoria de Exemplares e Fonologia de Uso. Adicionalmente, tal ferramenta é relevante para o desenvolvimento de pesquisas nas seguintes áreas: síntese de fala (Barnbrook, 2005), reconhecimento de fala (Barnbrook, 2005) processamento de linguagem natural (PLN) (Othero e Menuzzi, 2005), linguística de corpus (Sardinha, 2004), ensino e aprendizagem de língua materna e estrangeira (Hunston, 2002) e desenvolvimento de modelos teóricos que concebem a natureza probabilística da linguagem (Oakes, 2003;Bod, Hay e Jannedy, 2003). Em primeiro lugar serão apresentados os pressupostos teóricos que fundamentam a formulação e gerenciamento de um banco de dados de investigação segmental da fala. A busca por padrões segmentais específicos pode ser conjugada com o tipo silábico, o padrão acentual e a posição na palavra em que o segmento ocorre. Inicialmente será apresentada a primeira proposta de ferramenta de busca segmental que foi formulada em 2004 e concluída em 2005. Serão discutidos os sucessos e os problemas encontrados no gerenciamento do banco de dados e na pertinência das buscas realizadas. Em seguida será apresentada a nova versão da ferramenta que foi concluída em 2011, indicando-se as vantagens da segunda versão do buscador de parâmetros segmentais da linguagem. Será considerada uma avaliação complementar das duas versões do buscador segmental indicando-se as vantagens de cada uma das versões apresentadas. Em seguida, serão discutidos estudos de casos que indicam a pertinência e a adequação da proposta da incorporação de efeitos de frequência na investigação segmental da sonoridade da fala. Devido a necessidade de se evitar o uso de símbolos fonético as representações aqui apresentadas farão uso de símbolos alternativos aos símbolos fonéticos. Vários fenômenos fonológicos se encontram em curso, ou seja, há formas alternantes, em competição, em qualquer comunidade de fala. Por exemplo, em algumas variedades do português brasileiro temos a pronúncia dita ‘chiada’ no início da palavra ‘tia’ [tchia] como é o caso na variedade do português falado em Belo Horizonte. Já em outras regiões, como, por exemplo, em Recife, temos a pronúncia do som ‘t’ no início da palavra ‘tia’ [tia]. A alternância entre os sons [tch] e [t] representa o fenômeno de Palatalização de Oclusivas Alveolares, que engloba também as oclusivas alveolares vozeadas e as africadas alveopalatais vozeadas: [d] e [dj] como em [dia] e [djia] para a palavra ‘dia’. O fenômeno de Palatalização de Oclusivas Alveolares tem implicações em outros contextos da fonologia do

português brasileiro. Por exemplo, quando ocorre em uma palavra a sequência de sibilante e africada, ou seja [stch], por exemplo, na palavra 'vestido' [vestchidu] observa-se que a sibilante pode ser palatalizada, como em [veshtchidu], ou pode ocorrer somente uma fricativa [veshidu] (Guimarães, 2004). A alternância entre [stch]>[shtch]>[sh] abrange várias palavras que apresentam este padrão segmental. Por exemplo, as palavras: vestido, tristeza, justiça, diagnóstico, festival, etc. Observa-se que o padrão descrito aplica-se sistematicamente para os padrões segmentais [stch]>[shtch]>[sh]. Estes dados nos mostram que, em variedades lingüísticas diferentes, há padrões de sons em competição e que o padrão [stch]>[shtch]>[sh] é recorrente no português brasileiro. De acordo com dados do buscador segmental que apresentaremos um total de 2.015 palavras apresentam este tipo de padrão, contando com 1.616.461 ocorrências. Ou seja, este padrão sonoro é freqüente no português brasileiro. Contudo, para os correspondentes vozeados de tal padrão, ou seja [zdj]>[zjdj]>[zj], observamos que apenas 17 palavras foram encontradas no buscador, contando com 575 ocorrências. O fato de um padrão ser recorrente e produtivo, como no caso de [stch]>[shtch]>[sh], e de outro padrão ser restrito e improdutivo, como no caso de [zdj]>[zjdj]>[zj], oferece condições de implementarmos de maneira diferenciada tais padrões ao desenvolvermos estudos que tratem de tecnologia de fala (síntese e reconhecimento de fala), processamento de linguagem natural (PLN), linguística de corpus, implementação de ferramentas direcionadas para a educação (ensino de língua materna e estrangeira), implementação de técnicas de tratamento de saúde da fala (fonoaudiologia e clínica médica), além de contribuir com o desenvolvimento de modelos teóricos que concebam a natureza probabilística da linguagem. Os exemplos discutidos acima indicam a pertinência de se construir bancos de dados de informações de correlatos segmentais sonoros conjugados com informações de tonicidade, do tipo de sílaba e da posição do segmento na palavra. As informações sobre a sonoridade são conjugadas com informações de valores de freqüência em que os padrões sonoros ocorrem. Assim, o buscador segmental que apresentaremos consiste de uma ferramenta importante para o conhecimento de padrões sonoros do português brasileiro e pode contribuir com o conhecimento acurado da sonoridade quando aplicada a implementação de ferramentas tecnológicas nas áreas de síntese e reconhecimento de fala, processamento de linguagem natural, linguística de corpus, ensino e aprendizado de língua materna e estrangeira e desenvolvimento de modelos teóricos que concebem a natureza probabilística da linguagem.

Contato: thaiscristofaro@gmail.com

X Encontro de Linguística de Corpus
V Escola Brasileira de Linguística Computacional
9 a 12 de Novembro de 2011
Faculdade de Letras / UFMG

REALIZAR O ENQUADRAMENTO COMO EM FRAMES DE CONSTRUÇÕES COM VERBOS SUPORTE NO DOMÍNIO DO FUTEBOL: UM ESTUDO BASEADO EM CORPUS

Rove Chishman (Unisinos)
Guilherme Tiecher Figueiró (Unisinos)
João Gabriel Padilha (Unisinos)
Diego Spader (Unisinos)
Franciele Aguirres Pereira (Unisinos)

Este trabalho tem como objetivo apresentar um estudo semântico de sentenças com os verbos suporte dar e fazer presentes em um corpus voltado à temática do futebol em Português brasileiro. Esta investigação está ligada ao subprojeto Kicktionary_Br, que visa à construção de um corpus semanticamente anotado com base nas cenas e frames propostos pelo projeto pioneiro intitulado Kicktionary (Schmidt, 2009) – uma base de dados lexical online multilíngue que disponibiliza informação semântica sobre o léxico do futebol nas línguas inglesa, francesa e alemã com base em corpora. Essa plataforma online disponibiliza onze cenas que buscam dar conta dos momentos de uma partida de futebol, tais como shot (chute), pass (passe) e goal (gol). Essas cenas subdividem-se em frames, que representam suas diferentes possibilidades de realização em um jogo – por exemplo, a cena shot (chute) abrange os frames finish (finalizar), miss_Goal (perder_gol) e o frame homônimo shot (chute), entre alguns outros, que buscam dar conta das diferentes situações passíveis de realização quando a bola é chutada. Este estudo está relacionado ao projeto Framecorp, cuja meta principal é a anotação semântica de corpora por meio de frames para o Português brasileiro. A relevância desta investigação baseia-se, em um primeiro momento, na dificuldade de descrever semanticamente predicações constituídas por verbos-suporte – dar um chute, fazer (uma) falta, fazer (um) gol, dar (um) carrinho, etc. A descrição desse tipo de estrutura – constituída por um verbo associado a um substantivo em posição de objeto, de maneira a originar um sentido único a partir da união desses dois elementos – é dificultada pela notável irregularidade com que ocorrem na língua. Essa dificuldade de descrição relacionada às construções com verbos-suporte, por sua vez, incide sobre a metodologia de anotação semântica do corpus, uma vez que o arcabouço teórico da base de dados Kicktionary não prevê unidades lexicais relacionadas a esse tipo de estrutura. Basicamente, a evocação de frames é atribuída a verbos predicadores e substantivos. Por exemplo, é possível chegar-se ao frame shot de duas maneiras: (i) através do verbo chutar, e (ii) através do substantivo chute, não havendo hipótese para a construção com verbo-suporte dar um chute na lista de unidades lexicais referentes a esse frame. O estudo semântico dos verbos-suporte representa um desafio tanto para áreas como a lexicografia, quanto para sistemas que lidam com PLN (Processamento da Linguagem Natural), uma vez que a irregularidade dessas estruturas dificulta não apenas sua descrição do ponto de vista do significado, mas também sua padronização. Como consequência disso, emergem duas situações: primeiramente, dicionários eletrônicos compilados com base em corpora trazem muito pouco acerca do significado desse tipo de expressão, uma vez que elas possuem significação global, e não literal, inferida a partir da co-ocorrência de um verbo juntamente com seu complemento direto. Em um segundo momento, as construções com verbos suporte representam um problema para os tradutores automatizados, que tendem a interpretá-las literalmente, gerando significados ininteligíveis para os usuários desse tipo de ferramenta. A definição da metodologia a ser empregada para descrever semanticamente essas estruturas a partir dos dados

disponíveis no Kicktionary é uma das etapas fundamentais para a construção da versão brasileira desse recurso lexical voltado à temática do futebol. Os verbos-suporte – também conhecidos como verbalizadores – não se encaixam à categoria de predicadores típicos. Como postula Neves (2000), os verbos-suporte são aqueles verbos que possuem natureza semântica básica, e que, ao unirem-se a um sintagma nominal, passam a denotar funções morfológicas e sintáticas na construção do predicado. Essa combinação de dois elementos – um sintagma verbal e outro nominal – resulta em um significado único, que não pode ser inferido através da interpretação individual desses constituintes. Dar uma chance, dar um tempo, dar um grito, fazer uso, ter consideração, entre muitos outros, são exemplos de construções com verbos-suporte. A maioria dessas construções tem como correspondente semântico um verbo pleno, isto é, um verbo tipicamente predicador, o que permite paráfrases como: dar um tempo, esperar; dar um grito, gritar; fazer uso, usar; ter consideração, considerar, etc. Nessas construções, a função predicativa é transferida para o sintagma nominal, uma vez que o verbo-suporte conserva muito pouco de seu significado típico. Para a autora, diferentes razões motivam os falantes a optarem por esse tipo de estrutura: versatilidade sintática, adequação comunicativa – por exemplo, dar uma surra mostra-se menos formal do que surrar – e precisão semântica – tomar conhecimento, por exemplo, refere-se a um processo dinâmico, enquanto conhecer alude a um processo estático. Por ocorrerem de maneira pluriforme – fato que dificulta sua padronização – as construções com verbos-suporte, em estudos tradicionais como Katz (1975) e Makkai (1972), são tratadas como anomalias linguísticas, embora sejam linguisticamente muito produtivos. O corpus do projeto Kicktionary_Br é formado por 105 arquivos de texto que contêm notícias sobre os resultados de jogos de nove equipes brasileiras em três diferentes competições: Campeonato Brasileiro, Copa do Brasil e Copa Libertadores da América. Esses arquivos foram segmentados em 3.500 sentenças, as quais 397 envolvem construções com os verbos suporte dar e fazer, totalizando 73 e 309 ocorrências para cada um desses verbos, respectivamente. A extração dessas sentenças deu-se de maneira automática através do concordanceador Wordsmith. A partir da extração, foi realizada a verificação manual dos resultados encontrados a fim de eliminar as construções cristalizadas que utilizam tais verbos e as construções que não se referem aos eventos futebolísticos. Depois da exclusão de tais expressões, identificamos 43 e 275 ocorrências para os verbos dar e fazer respectivamente em que atuam como verbo suporte. Foi realizado o enquadramento das construções em suas respectivas cenas e frames e realizado um levantamento dos dados obtidos. Como resultado observamos que as construções com verbos suporte ocorrem principalmente nas cenas passe, chute, defesa, substituição e gol. A partir da análise desses dados evidenciamos que os sintagmas nominais que estão à direita dos referidos verbos suporte são os responsáveis pelo enquadramento da construção em determinado frame. Dessa forma, a etiquetagem das construções com verbo suporte ocorre de maneira distinta da habitual, uma vez que são os nominais os evocadores de frames.

TRABALHOS EM ANDAMENTO

UM PARALELO ENTRE O FRAME DE COMUNICAÇÃO DO PORTUGUÊS E DO INGLÊS

Francine Vaz (UFJF)

Essa pesquisa tem como fundamento a Semântica de frames, em especial os estudos relacionados a FrameNet objetivando a comparação dos frames da Língua Portuguesa e da Língua Inglesa no que tange o estudo do frame de Comunicação. Na Semântica de frames (FILLMORE, 1982), as palavras representam categorias de experiências, e cada uma dessas categorias é baseada por uma situação motivacional ocorrida em um background de conhecimento e experiência. A implementação mais amplamente desenvolvida dessa teoria é o projeto FrameNet (PETRUCK, 2009). O projeto FrameNet, em desenvolvimento desde 1997, é liderado pelo Professor Charles Fillmore no International Computer Science Institute (ICSI), em Berkeley, na Califórnia. De acordo com Ruppenhoffer et al (2006), o objetivo do projeto é criar "uma fonte lexical on-line baseada na semântica de frames e suportado por evidência de corpus". Dessa forma, torna-se possível documentar as possibilidades semânticas e sintáticas de cada palavra (valências) e de cada sentido dessa palavra através da anotação de frases exemplares e análise de resultados. Hoje, existem vários outros projetos que como a Framenet americana, procuram alcançar esses mesmos objetivos tendo outras línguas como foco, como por exemplo, o japonês, o espanhol, o alemão, o chinês e o português. Essa pesquisa está inserida no projeto FrameNet Brasil liderado pela professora Margarida Salomão na UFJF. No entanto, não é possível fazer uma transferência direta dos frames da Língua Inglesa para a Língua Portuguesa. Dessa forma, torna-se necessário fazer uma análise para saber até que ponto existe convergências e divergências. Seguindo as bases do projeto FrameNet, todas as análises devem ser baseadas em análise de corpus. Dessa forma, as sentenças utilizadas na análise de frame são retiradas de cinco corpora, sendo que os três primeiros são de domínio público disponíveis no site Linguateca (http://www.linguateca.pt/) e dois últimos de domínio privado disponíveis no site Sketch Engine (http://www.sketchengine.co.uk/): NILC/São Carlos, ANCIB, ECI-EBR, Nurc-RJ, Legendas de filmes. O foco desse trabalho é o frame de comunicação (um comunicador envia uma mensagem para um destinatário). Através do estudo do comportamento de verbos emblemáticos desse frame como comunicar, afirmar, informar, telefonar, faxear, telegrafar, prometer, permitir, discutir, gritar e sussurar no corpus é possível construir o frame de comunicação do Português com suas possíveis alterações. A anotação desses frames, assim como dos demais frames que fazem parte do projeto, permitirá o processamento computacional do português como inúmeras aplicações práticas no futuro, como a tradução automática entre línguas.

Contato: franfv@uol.com.br

O Vocabulário Do Horror: Uma Análise Contrastiva Bilíngue Baseada Em Corpus Do Léxico Especializado Da Série Supernatural

Raphael Marco Oliveira Carneiro (UFU/ILEEL)

A Linguística de Corpus tem proporcionado muitos avanços para a análise linguística por meio da compilação e da análise de corpora, que são constituídos por textos e transcrições de falas, armazenados em arquivos de computador (BERBER SARDINHA, 2009). Desse modo, neste trabalho foi possível analisar o léxico de uma das séries mais famosas e reconhecidas da atualidade: Supernatural. Entendendo que nessa série é apresentada uma área específica do saber (ocultismo) partimos da hipótese de que há um campo lexical especializado caracterizado pelo horror. Logo, podemos afirmar que há uma terminologia em Supernatural representativa de um conhecimento especializado. Por isso, na medida em que os termos apresentam tanto uma dimensão cognitiva, conceitual, quanto linguística, estes constituem o componente lexical especializado ou temático de uma língua (KRIEGER & FINATTO, 2004). Desse modo, o corpus estudado neste trabalho compõe-se das legendas em inglês e português das seis temporadas de Supernatural, sendo, pois, porções de falas transcritas produzidas por falantes nativos, e de suas traduções, realizadas por brasileiros, compreendidas em um período de tempo. Temos então uma amostra finita da linguagem como um todo, de conteúdo especializado, em que os textos são comparáveis paralelamente. A partir do objeto de pesquisa, assim caracterizado, os seguintes objetivos foram propostos: compilar e analisar o vocabulário específico proveniente do corpus bilíngue (Inglês/Português) das legendas da série Supernatural, bem como a elaboração de um glossário a partir da listagem dos termos. Os principais referenciais teóricos utilizados foram Berber Sardinha (2004, 2009), Fromm (2003, 2004, 2005) e Carvalho (2005) pelas suas respectivas contribuições na área da Linguística de Corpus, no uso de corpora na análise linguística e da tradução para legendas. Primeiramente, o corpus composto pelas legendas disponíveis on-line da série Supernatural foi compilado em inglês e em português. O uso da ferramenta WordSmith Tools versão 5.0 (SCOTT, 2008) permitiu o levantamento e a listagem das palavras-chave, bem como a listagem de concordâncias e a comparação das traduções, presentes no corpus. Desse modo, espera-se comprovar a predominância do vocabulário do horror, verificar como esse vocabulário foi traduzido e as prováveis adaptações resultantes do processo da tradução audiovisual para legendas.

Contato: raphael.olic@gmail.com

Referências:

BERBER SARDINHA, T. Linguística de Corpus. Barueri: Manole, 2004.

BERBER SARDINHA, T. Pesquisa em Linguística de Corpus com WordSmith Tools. Campinas: Mercado de Letras, 2009.

CARVALHO, C. A. de. A tradução para legendas: dos polissistemas à singularidade do tradutor. 2005. 160 f. Dissertação (Mestrado em Letras). Departamento de Letras do Centro de Teologia e

Ciências Humanas, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005. Disponível em:

<http://www.scribatraducoes.com.br/files/CarolinaAlfaroCarvalho_2005_TraducaoParaLegendas_Dissertacao.pdf>. Acesso em: 8 jan. 2011.

FINATTO, M. J. B. & KRIEGER, M. da G. Introdução à terminologia: teoria e prática. São Paulo: Contexto, 2004.

FROMM, G. A construção do sentido em vocabulários técnicos: o uso de corpora e outros procedimentos. Crop, São Paulo, v. 10, p. 225-239, 2005.

A Unidade De Apêndice De Comentário No Português Do Brasil

Cássia Jacqueline Fernandes Oliveira (PosLin / UFMG)

Este estudo se propõe a investigar a unidade de Apêndice de Comentário (APC), no Português do Brasil (PB). Foram analisados 20 textos de aproximadamente 1500 palavras. Esses textos compõem um minicorpus, formado com a mesma arquitetura do C-ORAL-BRASIL, corpus de fala espontânea, representativo da diatopia mineira. Os textos foram transcritos no formato CHILDESCLAN, implementado para a anotação informacional, e segmentados em enunciados, por meio do software WinPitch. Na amostra analisada encontramos 117 ocorrências de APC divididas entre os tipos conversação (36 ocorrências), diálogo (37 ocorrências) e monólogo (44 ocorrências). A análise proposta fundamenta-se na Teoria da Língua em Ato (CRESTI, 2000), cujo princípio baseia-se na correspondência entre unidade de ação e unidade linguística (enunciado), por meio de uma interface entonacional. O enunciado é entendido como a menor sequência linguística interpretável pragmaticamente, independentemente de uma composicionalidade sintática e de uma autonomia semântica. Para essa teoria, o APC é uma unidade de integração textual de uma unidade informativa de Comentário, com perfil nivelado ou descendente, em enunciados do tipo: “omitir /=COM= só //=APC”. Além desses três critérios, o APC pode ser classificado quanto ao seu caráter informacional (ELENA TUCCI, 2006), como: Repetições de expressões do tema (15%), Preenchimento (45%), Retomada textual (12%) e Informação tardia (33%). De um ponto de vista lexical a unidade de APC apresenta uma alta variabilidade morfossintática. Predominam-se os SN (20,5%), SP (20,4%), há os ADV (19,6%) seguidos pelos SV (17,1%), SADV (17%), Or. Sub. (4,2%) e SADJ (2,4%). Há mais três situações em que a unidade de APC pode ser realizada, embora em menor proporção que da ocorrência após a unidade de COM. São elas: 1) após unidades de Comentários Múltiplos (8,55%); 2) após unidades de Comentários Ligados (4,27%) e, por último, após unidades de Comentários Múltiplos Ligados (3,42%). Há, ainda, uma unidade que pode ocorrer em mesma distribuição do APC e gerar confusão, é a unidade de PAR (14,53%) (eles tavam num posto de gasolina /=COM= eu acho //=PAR=). Por se tratar de um estudo ainda em andamento, é nossa pretensão estabelecer a diferenciação, por todas as medidas, entre os textos monológicos, dialógicos e conversações.

Contato: cassiajfo@gmail.com

Referências:

CRESTI, E. (2000). Corpus di italiano parlato. Firenze: Accademia della Crusca, 2 voll.

MARTIN, P. WinPitch Corpus: A text to Speech Alignment Tool for Multimodal Corpora. Lisbon: LREC. May 2004. Disponível em: <http://lablita.dit.unifi.it/coralrom/papers/index.html>. Acesso em: 6 set 2007.

MELLO, H.; RASO, T. Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. Revista Portuguesa de Humanidades. 2009, pp. 301=325.

RASO, T.-MELLO, H. Parâmetros de compilação de um corpus oral: o caso do C-ORAL-BRASIL. In: Veredas.

TUCCI, E. (2005/2006). L'unità di appendice in un corpus di italiano parlato (C-ORAL-ROM): caracteristiche intonative, semantiche e morfo-sintattiche. Tesi de laurea triennale in italianistica. Universitá degli studi di Firenze, Facoltà di lettere e filosofia, anno academico.

O *DONDE* NO PORTUGUÊS E NO ESPANHOL: UM ESTUDO DE LINGUÍSTICA DE CORPUS

Alexia Duchowny (UFMG)
Simone Fonseca (UFMG)
Priscilla Santos (UFMG)

DONDE, tanto em português quanto em espanhol, é um termo que assume funções variadas, apresentando problemas tanto morfossintáticos quanto semânticos. Este trabalho propõe uma análise comparativa do DONDE no português e no espanhol, no século 15, para se entender melhor não apenas as diferenças entre estas duas línguas, mas também as semelhanças. Identificar as causas destes contrastes e destas afinidades, se genéticas ou tipológicas, por exemplo, também é uma questão de investigação. A metodologia de estudo selecionada é a da Linguística de corpus, representadas por autores como Kennedy (1998) e Sardinha (2004), caracterizada pela coleta e pela análise com critério e confiabilidade de corpora eletrônicos empregando-se ferramentas eletrônicas. Os corpora selecionados são os de M. Davies relativos ao português e ao espanhol, compostos de dados autênticos e bastante extensos (45 e 100 milhões de palavras, respectivamente). Por essas razões, são representativos das línguas em questão e permitem generalizações relativas ao funcionamento do português e do espanhol. Através do levantamento do DONDE em cada um dos corpora, no século 15, será possível, com os dois resultados preliminares, fazer-se uma comparação das duas línguas, o que permitiria ir ao encontro de informações que não podem ser encontradas quando se analisa cada uma das línguas individualmente. Espera-se, com esse estudo, compreender melhor esse termo de caráter polissêmico e polifuncional, para poder classificá-lo e defini-lo com mais precisão e detalhamento. Deseja-se, também, adquirir maior familiaridade com o programa dos dois corpora. Esta pesquisa insere-se em um trabalho mais amplo em andamento, que é o de elaboração de um quadro dos derivados de UBI e UNDE nas línguas românicas em geral.

Contato: alexiatelles@hotmail.com

Av. Antonio Carlos, 6627 Faculdade de Letras - UFMG - LEEL - Sala 4017

Referências:

DAVIES, M. Corpus del español. Disponível em: <http://www.corpusdelespanol.org/>. Acesso em: 18 mai. 2011.

DAVIES, M.; FERREIRA, M. O corpus do português. Disponível em:<http://www.corpusdoportugues.org/>. Acesso em: 18 mai. 2011.

KENNEDY, G. D. An introduction to Corpus Linguistics. Nova York: Longman, 1998.

SARDINHA, T. B. Lingüística de corpus. São Paulo: Manole, 2004.

LINGUÍSTICA DE CORPUS E TRADUÇÃO: SÉRIES AMERICANAS NO APRENDIZADO DE TERMINOGRAFIA BILÍNGUE

Guilherme Fromm (PPGEL/UFU)

Aprendizes de Tradução, ao serem apresentados à área de Terminologia Bilíngue, além da parte teórica, geralmente fazem um trabalho prático na elaboração de um glossário bilíngue, por exemplo. Muitos começam a trabalhar em subprojetos de seus professores/orientadores ou em áreas técnicas consagradas (como a jurídica, técnica/industrial, médica, entre outras) por eles indicados. Essa pesquisa indica uma nova abordagem para trabalho em sala de aula ou com alunos de Iniciação Científica: o uso de seriados de televisão para apreender os princípios básicos de um trabalho terminográfico. Com a atual facilidade de baixar seriados americanos na Internet (via Torrent), a tendência de vários desses seriados em trabalhar com uma terminologia específica, a elaboração de legendas em inglês por falantes nativos e sua subsequente tradução (sempre gratuitamente, realizada provavelmente por fãs não especializados na área de Tradução) e disponibilização em sites por brasileiros, os aprendizes têm um farto material disponível para o fazer terminográfico. Ao invés de trabalhar com textos jornalísticos, acadêmicos ou manuais sobre determinada área, o aprendiz descobre, através do seriado, como os termos dessa área são usados e como podem construir uma realidade mais plausível em obras de ficção. Com os princípios da Linguística de Corpus para compilação das legendas (elaboração de corpora escritos, sincrônicos/contemporâneos, por amostragem, bilíngues, de língua materna, paralelos e com a finalidade de estudo) e o uso de ferramentas de análise lexical, como o WordSmith Tools (aprendendo a trabalhar com as suas três ferramentas principais), o aluno pode levantar as palavras-chave do seriado na área da trama, tanto em inglês quanto em português, elaborar listagens dessas palavras-chave, construir definições e comparar as traduções feitas. Entre os vários exemplos de séries que trabalham com áreas de especialidade e que participam dessa pesquisa, podemos citar: House (médica), Law and Order (jurídica/criminal), CSI (criminal), Supernatural (ocultismo) e Farscape (astronáutica, astronomia, engenharia).

Contato: guifromm@ileel.ufu.br

A Tradução De Marcadores Culturais Na Relação De Cooperação Universitária Franco-Brasileira: Um Estudo Baseado Em Corpus

Aída Carla Rangel de Sousa (PGET/UFSC)

Este trabalho situa-se na área de estudos de tradução com base em córpus, no qual se relata a construção e análise de um córpus paralelo francês-português, no âmbito da cooperação universitária franco-brasileira. Admitindo que existam diferenças entre o Brasil e a França em áreas como educação, ensino e pesquisa, parece razoável pressupor que essas diferenças favorecem o aparecimento de termos não equivalentes nos níveis semântico e pragmático em textos dessas áreas. São exemplos do córpus, coletado a partir de sítios especializados bilíngues e multilíngues na Web, termos que remetem a realidades extralinguísticas específicas de cada sistema, como licença, bacharelado, mestrado, licence, master, dossier, grande école. Com o auxílio do programa WordSmith Tools (Scott, 2008), fez-se um levantamento de termos candidatos a marcadores culturais (MCs) e de padrões linguísticos relacionados a aspectos culturais de cada país envolvido, além de explicitar possíveis dificuldades de compreensão intercultural geradas por esses termos ou essas regularidades. Analisou-se a frequência desses MCs na lista de palavras-chave gerada na ferramenta KeyWords do WS Tools e verificou-se também a possível dicionarização desses MCs em ambas as línguas de estudo. Em seguida, empregou-se uma classificação de correspondências tradutórias (Thunes, 1998) que permitiu a análise das estratégias de tradução e suas implicações nos possíveis problemas de compreensão intercultural. Os quatro tipos de tradução propostos, hierarquizados de acordo com o grau de complexidade da informação, estão diretamente relacionados à forma de intervenção nos diferentes níveis linguísticos, sejam eles de natureza morfológica, sintática, semântica ou pragmática, para produzir o texto alvo. Embora não se trate de uma análise exaustiva, foi possível sugerir uma expansão do modelo de correspondências tradutórias proposto inicialmente, o qual está amparado no conhecimento das culturas envolvidas. Além disso, foi possível observar algumas características, motivadas pelas diferenças culturais, de um córpus no par francês-português de textos de especialidade.

Contato: aidacarlarangel@gmail.com

Referências:

SCOTT, M. WordSmith Tools version 5, Liverpool: Lexical Analysis Software, 2008.

THUNES, M. Classifying translational correspondences. In: Stig Johansson e SigneOksefjell (eds.). Corpora and cross-linguistic research: theory, method and case studies. Amsterdam: Rodopi, 1998, pp.25-50.

X Encontro de Linguística de Corpus
V Escola Brasileira de Linguística Computacional
9 a 12 de Novembro de 2011
Faculdade de Letras / UFMG

SUPERANDO O ESTADO DA ARTE NA ETIQUETAGEM MORFOSSINTÁTICA POR MEIO DE REGRAS DE PÓS-ETIQUETAGEM

Cid Ivan da Costa Carvalho (UFERSA)
Davis Macedo Vasconcelos (UFC/CEFET-CE)
Leonel Figueiredo de Alencar (UFC)

Segundo Kveton e Oliva (2002), os erros de etiquetagem, quando se utiliza um corpus para treinar estatisticamente algum sistema de processamento automático da linguagem natural, constituem desvios das regularidades que se espera que o sistema aprenda, resultando num modelo falso da língua. Eles propõem, desse modo, um método de correção desses erros num processo de pós-etiquetagem, levando em conta a detecção de n-gramas (em termos de seqüências de etiquetas) impossíveis na língua a ser modelada. Neste trabalho, procura-se sistematizar os erros cometidos pelo etiquetador morfossintático Aelius, construído por meio do NLTK (BIRD, KLEIN e LOPER, 2009). Este etiquetador, segundo Alencar (2010), foi treinado no Corpus Histórico do Português Tycho Brahe (CHPTB), utilizando em sua arquitetura o tagset desse corpus, que utiliza um sistema de anotação morfológica cujas etiquetas se subdividem em categoriais e flexionais. As primeiras classificam o item lexical em uma classe gramatical e as segundas indicam traços como gênero, número, pessoa, tempo e modo (CORPUS, 2010). Como o Aelius modela a morfossintaxe da língua portuguesa tal como se apresenta em textos literários dos séculos XV a XIX, período coberto pelo CHPTB, é natural que sua acurácia se reduza em textos contemporâneos, sobretudo quando se distanciam da língua padrão. Como ponto de partida deste trabalho, compilamos um corpus de textos de comunicação mediada por computador, gênero que se caracteriza pela abundância de abreviaturas, grafias não padrão e desvios da norma culta. Após a anotação feita com o Aelius, levantamos os erros cometidos pela ferramenta e fizemos uma correção manual. Em seguida, classificamos os erros cometidos e, com base nisso, elaboramos um primeiro conjunto de regras para correção automática da etiquetagem. Um segundo conjunto de regras procurou detectar e corrigir n-gramas impossíveis na língua portuguesa, conforme a proposta de Kveton e Oliva (2002). Numa segunda etapa do trabalho, tendo como meta chegar a 99% de acurácia, acima, portanto, do estado da arte de 97% (GÜNGÖR, 2010, p. 207), implementaremos essas regras em Python, a fim corrigir, num processo de pós-etiquetagem, os erros cometidos pelo Aelius na etiquetagem de textos atuais de comunicação mediada por computador.

Contato: cidivanc@ufersa.edu.br

Referências:

ALENCAR, L. F. de. Aelius: uma ferramenta para anotação automática de corpora usando o NLTK. ELC 2010 – IX Encontro de Linguística de Corpus, PUCRS, Porto Alegre, 8 e 9 de outubro de 2010. Disponível em: <http://corpuslg.org/gelc/elc2010.php> Acesso em: 28 fev. 2011.

BIRD, S.; KLEIN, E.; LOPER, E. Natural language processing with Python: analyzing text with the Natural Language Toolkit. Sebastopol: O'Reilly, 2009. 502 p.

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

GÜNGÖR, T. Part-of-Speech Tagging. In: INDURKHYA, N.; DAMERAU, F. J. Handbook of Natural Language Processing. 2. ed. Boca Raton, FL: Chapman & Hall/CRC, 2010. p. 205-235.

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

MODALIDADE EM CONSTRUÇÃO: UM ESTUDO COMPARATIVO BASEADO NO USO

Luciana Beatriz Bastos Avila (PosLin/UFMG)

Este trabalho tem como objetivo explorar o comportamento de construções, que têm como base verbos de crença como considerar, julgar, achar, crer, acreditar, supor, tomados como marcadores de modalidade. Pretendo, mais especificamente: (a) levantar hipóteses acerca da relação de um modelo teórico construcional e a modalidade; (b) investigar os contextos de uso das diferentes instâncias dessa construção; (c) analisar as funções discursivo-interacionais e os efeitos da relação entre os participantes; (d) discutir os conceitos de evidencialidade e modalidade. Assim, proponho que existe uma construção com os verbos elencados acima, que corresponderia a determinadas funções semântico-pragmáticas relacionadas à modalidade, analisada em contextos específicos de comunicação. Considero que um estudo sob um viés cognitivo (GOLDBERG, 2002, 2006) pode contribuir para se chegar a algumas respostas mais unificadas sobre a modalidade, por levar em conta diferentes dimensões: morfossintáticas, semânticas, pragmáticas. Tomo como base também as diretrizes metodológicas da Teoria da Língua em Ato (CRESTI, 2000), em que um enunciado é a unidade informacional mínima interpretada pragmaticamente, e que carrega um conteúdo ilocutório; a relação entre enunciado e ilocução é mediada pela prosódia, essencial, no âmbito desse trabalho, para a identificação dos índices modais. As ocorrências analisadas fazem parte de bases de dados da diamesia oral do português do Brasil e de Portugal. Utilizaremos dois subcorpora comparáveis, constituídos por vinte textos cada, entre monólogos, diálogos e conversações, da parte informal dos projetos C-ORAL-BRASIL (RASO; MELLO, 2009) e C-ORAL-ROM, braço português (CRESTI; MONEGLIA, 2005), distribuídos em contextos privados e públicos. Ainda, como referência, será usado o Corpus do Português (DAVIES; FERREIRA, 2006-). Para uma análise das listas de frequência, frequência relativa, concordanciador e contexto de uso será utilizado o software livre TextSTAT e as ferramentas contidas no próprio Corpus do Português. Como resultados, espero entender que: 1) há limites dos corpora de fala espontânea para análise de construções muito específicas; 2) a ocorrência das construções depende do propósito comunicativo e gêneros textuais específicos; 3) no PB e no PE, essas construções são usadas como uma avaliação do falante sobre a avaliação de uma terceira pessoa ou como proteção de face na interação.

Contato: lucianabeatrizavila@gmail.com

Referências:

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

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

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

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

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

RASO, T.; MELLO, H. Parâmetros de compilação de um corpus oral: o caso do C-ORAL-BRASIL. Veredas – Revista de Estudos Lingüísticos. Programa de Pós-Graduação em Lingüística. Juiz de Fora, 2009.

TextSTAT. Disponível em: http://neon.niederlandistik.fu-berlin.de/textstat/. Acesso em: 03 ago. 2010.

A Unidade De Introdutor Locutivo Em Um Corpus Oral De Português Brasileiro

Bruna Rocha (USP)
Bruno Rocha (PosLin/UFMG)

Neste trabalho são apresentados os primeiros resultados da análise da unidade informacional de introdutor locutivo (INT) no Português Brasileiro. Na Teoria da Língua em Ato (Cresti, 2000), base teórica desse trabalho, o introdutor locutivo é a unidade usada na fala para sinalizar que o espaço locutivo subseqüente apresenta um ponto de vista unitário que difere daquele do restante do texto. A Teoria da Língua em Ato parte da teoria dos Atos de Fala (Austin, 1962) e analisa a estrutura informacional de um enunciado levando em consideração a interface prosódica entre locução e ilocução. Para este estudo foi analisado um subcorpus composto de 20 textos de cerca de 1500 palavras cada (7 monólogos, 7 diálogos e 6 conversações), extraídos do C-ORAL-BRASIL, corpus de fala espontânea do Português Brasileiro representativo da diatopia mineira e portador da maior variação diafásica possível, com o objetivo de obter, assim, uma gama variada de ilocuções e estruturações informacionais. São descritas as características funcionais, distribucionais, prosódicas, morfossintáticas e lexicais do INT em PB. No subcorpus foram encontrados 231 INT, dos quais a maior parte (57%) encontra-se em monólogos. Do ponto de vista funcional, o INT introduz principalmente a meta-ilocução de discurso reportado (51%). Quanto à distribuição, localiza-se prioritariamente em posição inicial de enunciado (56%) e sempre imediatamente antes da unidade por ele introduzida. Quanto às características prosódicas, o INT é realizado mais rapidamente que o restante do enunciado e a sua freqüência fundamental (F0) é geralmente contrastante com a F0 da unidade seguinte. Estão sendo realizados testes com o objetivo de delimitar qual é o parâmetro mais relevante para a análise desse contraste (média de F0 ou pontos final e inicial do INT e da unidade subseqüente). Em relação às características sintagmáticas e morfossintáticas, os INT são constituídos principalmente de SV e seus verbos estão, na maior parte dos casos, no pretérito perfeito do indicativo. Futuramente esses resultados

serão comparados aos resultados obtidos para o INT em Italiano (Corsi, 2009) e poderão ser utilizados para a comparação interlingüística com outras línguas às quais já foi aplicada a mesma base teórica.

Contato: brunamaiarocha@gmail.com
bbruno791@yahoo.com.br

A Voz Do Tradutor Na Tradução Do Código Civil Brasileiro - Um Estudo De Tradução Baseado Em Corpus

Luciane Reiter Frohlich (PGET/UFSC)

A pesquisa aqui desenvolvida insere-se na nova realidade da pesquisa linguística eletronicamente assistida. Ela integra Estudos da Tradução e Linguística com base no estudo de corpora paralelo, com o auxílio do programa WordSmith, cujas ferramentas colaboram para o estabelecimento das relações entre os perfis micro (texto) e macro (cotexto) do corpus. O objetivo desta pesquisa é examinar empiricamente a voz do Tradutor em porções linguísticas extraídas da tradução de partes do Código Civil Brasileiro para a língua alemã, cujo corpus foi extraído de publicações bilíngues do portal advocatício Advokaturbüro Wolf/Suíça. Este objetivo está ligado ao escopo central deste estudo que visa ao levantamento de padrões linguísticos forenses e prováveis diferenças conceituais e culturais presentes entre as línguas envolvidas. Parte-se do pressuposto que a linguagem judicial apresente esses padrões e que eles representem de forma oficial marcadores culturais. Os textos traduzidos tratam exclusivamente de leis brasileiras, restringindo sua aplicação ao contexto legal brasileiro e/ou de brasileiros. Considerando este fato, é feito um levantamento dos recursos utilizados pelo tradutor (nota de rodapé, cursiva, parênteses, etc.) visando a exploração e demarcação de traços culturais ligados à linguagem forense no processo de passagem da língua portuguesa para a alemã, clarificando assim a voz do Tradutor. Conforme sugere Frankenberg-Garcia 2009, este trabalho passou por uma série de processos decisórios, desde sua definição até sua compilação e, como complemento, se propõe a seguir a sistematização metodológica proposta pela pesquisadora Lilian Fleuri (TraCor/PGET/UFSC 2011), em que a metodologia de construção, assim como a metodologia de análise do corpus, obedecem um padrão de organização e apresentação dos dados. Deste modo, espera-se contribuir para que os resultados obtidos não só permitam uma padronização da terminologia forense, mas dialoguem com outras investigações linguísticas semelhantes.

Contato: Luciane@i-trad.com

Referências:

BAKER, Mona (2004). A corpus-based view of similarity and difference in translation. In: International Journal of Corpus Linguistics 9:2. John Benjamins Publishing Company. p. 167-193.

FLEURI, Lilian (2011). Proposta de Sistematização Metodológica para Pesquisas em Análise Textual e Tradução: uma Interface com a Linguistica Sistêmico-Funcional e Os Estudos da Tradução Baseados em Corpus. In: I. Seminário TracCor. PGET-UFSC.

FRANKENBERG-GARCIA (2008). Compilação e uso de corpora paralelos. In: Avanços da Linguística de Corpus no Brasil. Stella Ortweiler Tagnin e Oto Araujo Vale (Orgs.). SP: Humanitas. p. 117-134.

SARDINHA, Tony Berber (2004). Linguística de corpus. Barueri, SP: Manole.

Referência online: http://www.law-wolf.ch/gesetze-de.htm

CORPUS DE GÊNEROS ACADÊMICOS DE LINGUÍSTICA E LITERATURA: COMPILAÇÃO, ANÁLISE E APLICAÇÕES DIDÁTICAS

Simone Sarmento (UFRGS)
Bruno Scortegagna (UFRGS)
Larissa Goulart da Silva (UFRGS)

A presente pesquisa tem por objetivo compilar um corpus de gêneros acadêmicos da área de Linguística e Literatura, nas línguas inglesa e portuguesa, de forma a possibilitar a análise linguística e a elaboração de projetos e tarefas pedagógicas com foco no aluno do Instituto de Letras (IL) da Universidade Federal do Rio Grande do Sul (UFRGS), futuro professor de línguas e/ou tradutor. O estudo foi motivado a partir da percepção compartilhada entre os professores de língua inglesa do IL da UFRGS de que há uma carência de materiais didáticos baseados em textos autênticos para o ensino de Inglês Acadêmico para essa área. A primeira parte da pesquisa dedicar-se-á a compilação do gênero abstract. O corpus de abstracts de língua inglesa e portuguesa está subdividido em 14 subcorpora: abstracts de TCCs, teses e dissertações (Ing/Port-Linguística e Literatura) provenientes dos alunos do IL da UFRGS, abstracts de revistas nacionais de Literatura e Linguística QualisA (Ing/Port) e revistas internacionais de Literatura e Linguística QualisA (Inglês). Através dessa subdivisão será possível verificar quais as áreas linguísticas mais prementes e importantes de serem abordadas nos materiais didáticos. Esse corpus será disponibilizado na base de dados do projeto TERMISUL (http://www6.ufrgs.br/termisul/ disponível a partir de agosto/2011). Após a compilação, será realizado o estudo do corpus com o auxílio do software WordSmith Tools v.5. Nesta fase, serão utilizados diferentes recursos do software, como lista de palavras, concordâncias, lista de colocados, entre outros. Posteriormente, a partir dos resultados dessa investigação, será realizada a elaboração de tarefas pedagógicas para o ensino da leitura e redação de abstracts em língua inglesa. Como a pesquisa ainda está em fase inicial, apresentaremos resultados parciais referentes às escolhas gramaticais e lexicais feitas pelos diferentes grupos em análise, ou seja, alunos de graduação e pós-graduação do curso de Letras e pesquisadores experientes.

Contato: Simone_sarmento@terra.com.br

Referências:

ESTAIRE, S. e ZANNON, J. 1994. Planning classroom: A task-based approach. Oxford, Heinemann.

BERBER SARDINHA, A.P. Lingüística de Corpus. Barueri: Manole, 2004.

BERBER SARDINHA, A.P. Preparação de material didático para aprendizagem baseada em tarefas com WordSmith Tools e corpora. In Calidoscópio, vol.4, n.3, p.148-155, set/dez 2006.

SARMENTO, S. O uso dos verbos modais em manuais de aviação em inglês: Um estudo baseado em corpus. Tese de doutorado. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2008.

VIANA, V.; TAGNIN, S. E. O. Corpora no ensino de línguas estrangeiras. São Paulo:

Hub Editorial Ltda., 2011.

FORMAS MODAIS EQUIVALENTES COM VALORES SEMÂNTICOS DIVERSOS: UM MAPEAMENTO EM CORPORA

Raíssa Caetano (IC/UFMG)
Luis Lima e Silva (IC/UFMG)

O presente trabalho desenvolve-se em dois momentos: um, inicial, em que se objetivou observar a ocorrência de duas variantes de marcadores modais, uma adverbial e outra predicativa, e verificar se estas possuíam o mesmo valor semântico em contextos específicos; e outro que visa discutir hipóteses acerca dos resultados da primeira parte a partir dos usos característicos das formas em tela. Os pares de itens investigados são: realmente/na realidade, obviamente/é óbvio, claramente/é claro e logicamente/é lógico. Não trataremos de questões intrínsecas à discussão da categoria modalidade, visto que há uma gama de estudos, muitas vezes controversos, sobre o fenômeno. A asserção que motivou a discussão do fenômeno considerado foi a de Halliday (1970), em que este menciona a não-equivalência de formas como certamente/tenho certeza de que, desenvolvendo mais o tema. Essa mesma consideração de Halliday (1970) é discutida por García (2000) ao comentar estudos sobre modalidade. O autor ressalta a recorrência de inúmeras "combinações" de marcadores em uma mesma parte do discurso, ou seja, uma possibilidade de diferentes formas de distribuição estrutural na cláusula. Nesta pesquisa, especial atenção será dada à metodologia; através da comparação entre corpora pretende-se mapear a ocorrência dos itens. Os corpora utilizados na primeira parte são: um corpus do Português (Davies e Ferreira, 2006-), tomado como de referência, e um corpus de fala espontânea do PB denominado C-ORAL-BRASIL (Raso e Mello, 2010). A partir da análise de ocorrências nos dois corpora, obtivemos alguns resultados e fomos guiados ao levantamento de novas hipóteses. As considerações feitas perante os dados apontaram a necessidade de discussão do estatuto do item realmente no PB e, ainda, fizeram com que se pensasse no estudo de uma "prosódia semântica" em novas investigações. Com relação ao advérbio realmente, tornou-se necessário a investigação diacrônica do item e seus possíveis correlatos, tarefa que está sendo executada no corpus de referência do português. No que diz respeito aos demais itens em análise, além da prosódia semântica, há que se mencionar outras noções relacionadas ao campo semântico da modalidade, a saber, graus de certeza e polidez.

Contato: raissavoliveira@gmail.com

X Encontro de Linguística de Corpus
V Escola Brasileira de Linguística Computacional
9 a 12 de Novembro de 2011
Faculdade de Letras / UFMG

PROBLEMAS NA CONSTITUIÇÃO DE UM CORPUS PARA UM DICIONÁRIO DE APRENDIZES DE ESPANHOL

Carolina Alves (UFRGS)
Carolina Reolon Jardim (UFRGS)
Laura Campos de Borba (UFRGS)
Félix Bugueño Miranda (UFRGS)

A Linguística de Corpus apresenta uma evidente interface com a Lexicografia. Antes de seu surgimento no cenário dos estudos da linguagem, a compilação de corpora para fins de elaboração de dicionários era realizada manualmente. Hoje, com o auxílio de ferramentas computacionais, tal processo é quase todo automatizado. Contudo, percorrer o caminho entre a compilação de um corpus e o dicionário como produto final, não é tão simples quanto parece. Existe uma ampla gama de considerações a serem feitas, por um lado, no que se refere à parte de coleta de dados, e, por outro, no que condiz à Lexicografia. O presente trabalho integra um dos módulos de um projeto maior que pretende estabelecer as bases teórico-metodológicas para a elaboração de um dicionário monolíngue de espanhol como LE para aprendizes brasileiros. No atual estágio de nossa pesquisa, começamos a compilar um corpus a fim de estabelecer como deve fundamentar-se a marcação diatópica das unidades léxicas do espanhol na obra lexicográfica a que se destina nosso estudo. Portanto, o objetivo da presente comunicação é apresentar uma reflexão referente a aspectos teóricos e metodológicos do trabalho desenvolvido até agora. Como metodologia, analisaremos os dados obtidos à luz de critérios intra e extralinguísticos referentes à diferenciação diatópica entre as variantes da língua espanhola, cotejando tais dados com os pressupostos da Linguística de Corpus. Nossos primeiros resultados corroboram com nossa hipótese de que a viabilidade da constituição de um corpus depende não somente de problemas inerentes à dialetologia, mas também dos resultados efetivos que se possam obter com o uso de softwares utilizados para compilar corpora.

Contato: carolespanhol@gmail.com

Referências:

FRAGO GRACIA, J.A.; FRANCO FIGUEROA, M. El español de América. Cádiz: UCA, 2003.

LIPSKI, J.M. El español de América. Madrid: Cátedra, 1996.

MORENO DE ALBA, J. G. El español en América. México, D.F.: FCE, 1993.

SARDINHA, T.B. Linguística de Corpus. Barueri: Manole, 2004.

ANALISANDO OS ITENS MAIS FREQUENTES EM UM CORPUS ORAL DE APRENDIZES DE LÍNGUA INGLESA: DESCOBERTAS E REFLEXÕES

Thaís Helena Pereira Marques (UFSJ)
Bárbara Malveira Orfanò (UFSJ)

Embora os estudos baseados na interlíngua de aprendizes da língua inglesa ainda sejam escassos, Dutra e Silero (2010) observam que a área da interlíngua tem tido um papel importante para a compreensão do que é a produção de aprendizes. Nessa perspectiva, este trabalho tem como objetivo analisar o discurso oral de um grupo de alunos da graduação do curso de Letras (Licenciatura em Inglês) da Universidade Federal de São João del-Rei, usando ferramentas da linguística de corpus (palavras chaves, listas de frequência, linhas de concordância e clusters). Essa pesquisa propõe analisar os itens lexicais mais frequentes em um corpus de aprendizes para que se possam identificar aspectos importantes no discurso espontâneo dos alunos. Para tanto, está sendo compilado um corpus oral de pequena dimensão (Berber-Sardinha, 2004) coletado durante as aulas da disciplina Conversação em Língua Inglesa, ministrada no primeiro semestre de 2011, que será contrastado com um sub-corpus de falantes nativos (The Santa Barbara Corpus of Spoken Language). Resultados preliminares apontam diferenças interessantes tanto quanto na forma e no uso dos bundles em cada corpus. Por exemplo, o sobreuso da expressão and things no discurso dos alunos brasileiros indica que este bundle exerce uma função diferenciada no corpus de aprendizes. Acredita-se que a identificação dos diferentes tipos de bundles nos dois corpora evidencia aspectos importantes da interlíngua do corpus de aprendizes desse estudo, que podem vir a contribuir para o ensino e aprendizagem das habilidades orais em língua inglesa dos alunos em questão.

Contato: thaisletrasufsj@gmail.com

O USO DE CHUNKS FORMADOS PELO VERBO GET POR APRENDIZES DE INGLÊS COMO L2

Gláucio Fernandes (PosLin/UFMG)

Os chunks como um tipo de regularidade da língua no nível da forma e do significado, têm ganhado considerável atenção nos últimos anos nos estudos linguísticos em geral e no campo de ensino/aprendizagem de línguas em particular. Neste trabalho, buscamos observar os caminhos que levam à aquisição, compreensão, e produção de chunks em língua inglesa. Para isso, revisaremos trabalhos anteriores que tratam essas mesmas construções, denominando-as 'verbos frasais'. Posteriormente, nos embasaremos em teóricos que tratam a diferença entre línguas verb-framed e línguas satellite-framed assim como aqueles que discutem a hipótese da transferência de língua. Esse trabalho busca investigar o uso de chunks com o verbo get, contrastando o seu papel no inglês L1 e o dos seus correspondentes em português L1, com fins à compreensão do seu uso no inglês L2 de falantes nativos de português L1. O objetivo central desta pesquisa é identificar o uso de chunks com o verbo get em construções de movimento por aprendizes brasileiros de inglês L2. Metodologicamente, o trabalho segue a orientação da Linguística de Corpus. A partir da produção linguística desenvolvida por 50 participantes em duas tarefas propostas, observaremos a

capacidade de uso de chunks formados pelo verbo get + Path, por aprendizes de inglês como L2. Os resultados encontrados nos possibilitará fazer uma análise comparativa das formas e suas frequências com aquelas encontradas nas narrativas colhidas em língua inglesa L1 e em língua portuguesa L1. Dessa maneira, a partir da coleta desse minicorpus, buscaremos analisar os chunks usados pelos participantes, mostrando a distribuição de frequência do verbo get em chunks produzidos por aprendizes de inglês L2. Isso poderá ser feito pois a transcrição dos dados obtidos através da produção dos aprendizes será analisada por meio do software TextSTAT. Após o tratamento de freqüência, os dados serão analisados estatisticamente para se observar a associação entre pares de eventos. A análise dos dados se dará de maneira que seja possível observar a convergência e a transferência das estruturas (chunks) formadas a partir do verbo get no inglês e a correlação com verbos do português.

Contato: glauciocalama@yahoo.com.br

A Elaboração De Ontologias Linguísticas No Âmbito Da Web Semântica: Uma Ontologia No Domínio Da Indústria De Artefatos De Borracha

Abner Maicon Fortunato Batista (UNESP)
Claudia Zavaglia (UNESP)

Atualmente, ontologias têm ganhado destaque por viabilizarem importantes aplicações computacionais. Uma delas, que vem ganhando destaque nos últimos anos, é a de um novo conceito de Web denominado Web Semântica, em que a informação possui um significado claro, bem definido e estruturado, possibilitando uma melhor interação entre computadores e pessoas (BERNERS-LEE, 2001). Este trabalho tem como objetivo geral realizar um estudo sobre a elaboração de ontologias procurando dar ênfase à relevância de teorias e recursos linguísticos, como corpus, na construção de ontologias para a Web Semântica. A partir de uma pesquisa realizada por Bazzon (2009) na Universidade Federal de São Carlos a respeito da criação de um vocabulário no domínio da Indústria de Artefatos de Borracha– IAB, este trabalho tem como objetivo específico elaborar uma ontologia para esse domínio. A questão que se coloca é se o uso de uma abordagem linguística, utilizando evidências empíricas contidas em corpus pode ser eficiente para a elaboração de ontologias no âmbito da Web Semântica. Os entornos semânticos do referido domínio são observados na busca de equivalências no corpus de IAB (contendo 1.200.000 palavras). Os conceitos são organizados a partir da Estrutura Qualia do Léxico Gerativo de James Pustejovsky (1995), que especifica quatro aspectos essenciais do sentido de uma palavra: constitutivo, formal, télico e agentivo. Em seguida, o esquema conceitual deverá ser implementado em OWL (Ontology Web Language), uma linguagem computacional voltada para Web Semântica. Finalmente, pretendemos realizar os testes e refinamentos na ontologia para resolver quaisquer problemas decorrentes do processo. Concomitantemente a todas as etapas previstas, é gerada uma documentação dos processos para viabilizar o posterior reuso da ontologia em aplicações diversas, bem como sua integração com outras ontologias existentes. Até o momento, foram extraídos e validados 130 termos, que foram classificados em quatro subdomínios: matéria-prima, equipamentos, processos e produto final. Os resultados esperados

com este trabalho são a elaboração da ontologia no domínio de IAB e a sua implementação computacional para uso em Web Semântica.

Contato: abnerfortunato@gmail.com

Referências:

BAZZON, S. C. M. Terminologia da indústria de artefatos de borracha: proposta de um vocabulário. Dissertação de mestrado. Universidade Federal de São Carlos – UFSCar. São Carlos, 2009.

BERNERS-LEE, T., HENDLER, J.; LASSILA, O., The Semantic Web, Scientific American, 284, pp.34-43, 2001.

PUSTEJOVSKY, J. The Generative Lexicon. Cambridge: The MIT Press, 1995.

SOBRE A EXPRESSÃO LINGUÍSTICA DE EVENTOS DE TRANSFERÊNCIA DE POSSE NA INTERLÍNGUA PORTUGUÊS-INGLÊS

Júlia Zara (PosLin/UFMG)

A partir da consideração de que a proficiência em uma língua envolve não apenas o conhecimento das construções que desta fazem parte, mas também a capacidade de selecionar estas construções em seus contextos preferenciais de uso (WULFF & GRIES, 2006), este trabalho investiga o uso de construções da estrutura argumental na expressão linguística de eventos de transferência de posse por falantes da interlíngua português-inglês em diferentes níveis de proficiência na língua-alvo. Para alcançar resultados que sejam estatisticamente confiáveis e generalisáveis acerca do uso das construções, serão utilizados recursos da Linguística de Corpus. Primeiramente, corpora do inglês e do português já existentes serão analisados utilizando-se o software R (GRIES, 2009). Nesta parte da pesquisa, serão calculados o número total de ocorrências de cada construção nos corpora e a frequência de ocorrência de cada construção com as seguintes variáveis: verbo utilizado, propriedades dos argumentos <tema> e <recipiente> e canal. Os resultados serão submetidos a testes de significância estatística (McEnery & Wilson, 2001). A partir da comparação entre os resultados obtidos para o português e para o inglês, serão elaboradas tarefas experimentais através das quais dados escritos e orais da interlíngua serão coletados para a formação de um corpus experimental, ideal para se testar a influência de fatores específicos no uso de construções linguísticas (Gilquin & Gries, 2009). Uma análise qualitativa da interlíngua permitirá uma descrição rica e detalhada sobre os padrões deste sistema linguístico, os quais serão, em seguida, analisados quantitativamente. Finalmente, os resultados obtidos para os três sistemas linguísticos serão comparados. Apesar dos corpora possuirem tamanhos distintos, existem testes de significância estatísticas através dos quais é possível estabelecer este tipo de comparação (McEnery & Wilson, 2001). Assim, será possível identificar: possíveis influências da português na interlíngua (ODLIN, 1989), padrões da interlíngua que se diferenciam tanto do português quanto da inglês e padrões da interlíngua que se aproximam do inglês. Espera-se observar transferência seletiva de elementos do português para a interlíngua nos estágios iniciais

de desenvolvimento e crescente aproximação aos padrões do inglês em níveis de proficiência mais avançados.

Contato: juliavidigal@yahoo.com

ESTUDO DO "NÃO" REFERENTE A NOMES NA HISTÓRIA DO PORTUGUÊS: CONSTITUIÇÃO DE UM CORPUS DIACRÔNICO A PARTIR DO BANCO DE DADOS DE DAVIES E FERREIRA (2006)

Pâmella Alves Pereira (IC/UFMG)

Este estudo, que integra o grupo de pesquisa Varfon-minas – variação lexical, morfológica e fonético-fonológica em Minas, tem como objetivo a análise do elemento "não" quando anteposto a nomes (doravante formações do tipo "não + nome"), especificamente, a palavras de valor adjetival e substantivos. Definir o estatuto gramatical do "não" nesse tipo de formação não é simples, já que nem todos compartilham a mesma opinião quanto à classificação de "não sócio", por exemplo: ora é classificado como um composto morfológico, formado por duas bases ("não" + "sócio"), ora como uma palavra formada a partir de um prefixo anteposto a uma base ("não-" + "sócio") e há, ainda, quem considera que esse "não" anteposto a nomes não é parte de um composto morfológico nem é um prefixo, conforme apresentam Silva & Mioto (2009). Em nossa análise temos como suporte teórico, principalmente, a teoria da Gramaticalização (Hopper & Traugott, 1993), por isso foi necessário a constituição de um corpus com dados históricos do Português, já que entendemos a gramaticalização como um processo diacrônico. Nesse sentido, foram coletados dados dos séculos XIV ao XX do banco de dados de Davies & Ferreira (2006-), denominado "O Corpus do Português" (doravante CdP) e disponível em http://wwwcorpusdoportugues.org. Utilizamos essa ferramenta eletrônica para a coleta de dados por ela permitir uma pesquisa rápida em mais de cinquenta mil textos do Português de diferentes épocas. Para realizarmos nossa coleta dos dados no "Corpus do Português" (Davies & Ferreira, 2006-) utilizamos a seguinte sintaxe de consulta: não *, que significa que procuramos por todos os casos de "não" seguido por qualquer palavra, sendo que deve haver um espaço em branco entre o "não" e a palavra seguinte. Ao fazermos essa busca em cada século, o CdP ofereceu-nos como resultado um lista de ocorrências, entre elas, todos os casos de "não" anteposto a verbo, pronome, adjetivo, substantivo, enfim, todas as classes gramaticais possíveis. Para nossa análise interessam apenas os dados em que o "não" antepõe-se a um particípio com valor adjetival, a um adjetivo ou a um substantivo, por isso foi necessário ler cada dado e selecionar apenas aqueles que integrariam nosso corpus. Essa primeira coleta resultou em dados de formações do tipo "não + nome" sem hífen. Para finalizá-la, precisávamos incluir os casos com hífen e, para isso, pesquisamos no CdP a seguinte sintaxe: não*, que significa que pretendíamos encontrar todos os casos de "não" seguido por qualquer coisa sem o espaço em branco. Essa pesquisa teve como resultado, em sua maioria, casos de "não" seguido imediatamente por hífen, e o hífen imediatamente seguido por outra palavra. Selecionamos os dados e constituímos, assim, nosso corpus. Como estamos diante de uma análise histórica, é certo que a palavra "não" apresente grafias diferentes e precisávamos considerar cada uma na constituição do nosso banco de dados. O CdP disponibiliza dados de textos antigos do Português, e encontramos, principalmente nos séculos XIV a XVIII, as seguintes grafias da palavra "não": nom, non, nam, nan, nõ, nã, naõ, nao (sem o til) e não. Fizemos, assim, uma pesquisa para cada uma dessas possibilidades de grafia.

Após a coleta e a seleção dos dados, constituímos um corpus do Português com 6761 casos de "não" anteposto a nomes, do século XIV ao XX. Os dados foram organizados segundo diferentes critérios: primeiramente, separamos os casos com hífen dos casos sem hífen, depois dividimos os casos de "não + nome" em: "não" referente a um particípio; "não" referente a um adjetivo e "não" referente a um substantivo, por último organizamos cada caso conforme critérios sintáticos e semânticos. Verificamos a frequência de ocorrência de "não + nome" no corpus organizado e analisamos a trajetória desse tipo de formação ao longo da história da língua. É importante dizer ainda, que, entre os dados de "não" seguido por um adjetivo, há os casos da expressão "não obstante", que foram coletados, organizados e analisados separadamente. Os resultados obtidos até então mostraram que as formações em estudo não são tão recentes como se pensava, pois encontramos dados de "não" anteposto a palavras adjetivais desde o século XIV, e "não" anteposto a substantivos desde o século XV. De fato, esse tipo de formação mostrou-se mais produtivo no século XX, e as formações de "não" seguido por um substantivo foram a que tiveram aumento mais significativo nesse século. Sobre os casos de "não obstante", especificamente, verificamos em sua trajetória ao longo da história da língua indícios de que a expressão, hoje entendida como uma locução conjuntiva com valor concessivo ou adversativo, tenha passado por um processo de gramaticalização.

Contato: pammelaalvespereira@gmail.com

Referências:

DAVIES, M. & FERREIRA, M. J. O corpus do Português [online] Disponível em: http://corpusdoportugues.org. 2006-

HOPPER, P. & E. TRAUGOTT. Gramaticalization. Cambridge: Cambridge University Press. 1993

SILVA, M. C. F & MIOTO, C. Considerações sobre a prefixação. ReVEL, vol. 7, n. 12. [www.revel.inf.br]. 2009

LINGUISTICA DE CORPUS – MAIS UMA POSSIBILIDADE NO ENSINO DE LÍNGUA ESTRANGEIRA PARA ALUNOS COM NECESSIDADES ESPECÍFICAS

Marlene Andreetto (USP)

O objetivo desta pesquisa é mostrar, fazendo uso da LC, uma possibilidade de promover autonomia no aprendizado de uma língua estrangeira. Segundo Vygotsky (1996), não se deve saber o que o aprendiz pode fazer no momento e sim o que ele será capaz de fazer no futuro com seus próprios recursos, através de uma mediação adequada; que pode vir a partir de um instrutor ou mesmo um instrumento, no caso o computador. A pesquisa surgiu da necessidade de preparar alunos para exames de proficiência em leitura de textos acadêmicos. A principal questão a ser investigada é saber a razão pela qual, alunos costumam encontrar maior dificuldade em textos de divulgação científica e não textos acadêmicos de suas áreas de especialidade. Os exames, acima referidos, são compostos de textos mais de divulgação do que científicos propriamente ditos. Para isso estamos, estamos coletando um corpus em inglês composto de textos acadêmicos, textos de

notícias do site BBC News e algumas amostras de textos de exames anteriores Pretendemos verificar porque esses alunos consideram textos de divulgação mais difíceis do que específicos da sua área de especialidade. Seriam as estruturas, o vocabulário? Como ponto de partida de investigação, colocamos três exemplares de cada gênero textual no site do Lextutor (o número de textos se deve a limitação de caracteres da ferramenta) e verificamos que a densidade lexical é um dos pontos importantes a serem investigados . Além disso, verificamos que a porcentagem das palavras cognatas nos textos mais acadêmicos é bastante alta, o que parecer indicar uma relação entre o foco da nossa investigação e o problema encontrado pelos aprendizes de leitura. Como parte da investigação, estamos distribuindo aos alunos, na sua maioria nível pré- intermediário em inglês três diferente tipos de textos (acadêmicos, de notícias e presentes nos exames) para que eles possam identificar e nos informar quais as dificuldades encontradas em cada um e na percepção deles, quais seriam os pontos vulneráveis. Ao mesmo tempo, estamos fazendo listas de palavras de palavras chave dos diferentes sub-corpora, utilizando o Wordsmith Tools (Scott, 2009), já que essa ferramenta a nos proporciona melhores condições de armazenamento e além de recursos apropriados para uma análise mais detalhada. Dessa forma, esperamos poder chegar a resultados que possibilitem uma construção de saberes conjunta entre pesquisadores, professores e aprendizes.

Contato: maandreeto@yahoo.com

Referências:

Berber Sardinha, T. Linguística de Corpus. Barueri , SP: Manole, 2004.

Scott, M. WordSmith Tools. Versão 5. Oxford: Oxford University Press, 2009.

Scott, M.; Tribble, Textual paterns: key words and corpus analysis in language education. Amesterdam: John Benjamins, 2006.

Viana, V e Tagnin, Stella E. O. Corpora no Ensino de Línguas Estrangeiras . Hub Editorial, 2011.

Vygotsky, L. S. A Formação Social da Mente. São Paulo: Martins Fontes, 1996.

http://www.lextutor.ca/vp/eng/

http://www. bbcnews.co.uk

A Linguística De Corpus Na Elaboração De Tarefas Pedagógicas Para Graduandos Em Letras-Inglês

Ana Paula Seixas Vial (IC/UFRGS)
Anamaria Welp (UFRGS)

O presente trabalho apresenta um projeto em andamento na UFRGS que visa à elaboração de material didático voltado a estudantes das áreas de ensino e tradução de inglês. Para tanto, foram integradas três áreas de pesquisa, a Linguística de Corpus, a Análise de Gêneros e o Ensino por Tarefas. O objetivo do trabalho é a criação de modelos (templates) de dois tipos de tarefas com

base em Berber Sardinha (2011). A primeira tarefa é centrada na concordância e tem o objetivo de tornar alunos pesquisadores, treinando-os a buscar nas concordâncias regularidades no uso autêntico da língua e a notar nessas regularidades questões importantes, como o significado de palavras e expressões ou o uso de classes gramaticais, através de atividades guiadas, contendo perguntas e exercícios voltados para aspectos específicos da concordância. Para sua execução, pretende-se recorrer a compilações já existentes, como o British National Corpus (BNC) e o Corpus of Contemporary American English (COCA). A segunda tarefa consistirá em perguntas e exercícios divididos em quatro partes: foco na atividade social, foco no gênero, foco no texto e foco no corpus. Essa última parte propõe uma atividade de exploração deum corpus do gênero estudado na disciplina a qual a tarefa é destinada. Como o corpus do gênero específico não precisa ser extenso desde que seja suficiente para revelar aspectos sobre sua linguagem, será construído um corpus de algumas dezenas de textos, colhidos da Internet. As perguntas e os exercícios contidos na tarefa serão padronizados, aplicando-se a qualquer outro gênero escolhido. Espera-se, com os resultados, atrair interesse de futuros profissionais da língua para a Linguística de Corpus e disponibilizar para o Instituto de Letras da UFRGS e para a comunidade acadêmica um material especialmente elaborado para o curso.

Contato: anapvial@gmail.com

VERBNET.BR: CONSTRUÇÃO SEMIAUTOMÁTICA DE UM LÉXICO VERBAL ONLINE E INDEPENDENTE DE DOMÍNIO PARA O PORTUGUÊS DO BRASIL

Carolina Scarton (USP-São Carlos)
Sandra Aluisio (USP-São Carlos)

Dentre as atividades compreendidas pela área de Processamento de Línguas Naturais encontram-se a criação e a disponibilização de recursos léxicos computacionais (RLC's). Neste trabalho, é definido um método de construção semiautomática (em quatro etapas) de um RLC de verbos para o português do Brasil, chamado VerbNet.Br, que segue os moldes da VerbNet (Kipper, 2005) que, por sua vez, segue a teoria das classes de Levin (1993). O trabalho aqui descrito é baseado em duas hipóteses: é possível a criação de um recurso léxico para o português de mesmas características da VerbNet, diretamente alinhado com a mesma, via um método semiautomático que se utiliza de recursos existentes, aproveitando-se do potencial cross-linguístico das classes de Levin (Jackendoff, 1990); e o método semiatomático produzirá resultados mais precisos do que métodos totalmente automáticos.. A primeira etapa (manual) consiste da tradução das alternâncias sintáticas da VerbNet que podem ser diretamente traduzidas para o português. A segunda etapa (automática) pretende utilizar a Linguística de Corpus para extrair informação das alternâncias sintáticas das quais os verbos no português participam. A terceira etapa (automática), já implementada, utilizou recursos léxicos computacionais já existentes (WordNet.Br) para a definição dos candidatos a verbos membros das classes VerbNet.Br. Por fim, na etapa final (automática) para cada verbo candidato serão comparadas as alternâncias definidas manualmente para a classe com as alternâncias encontradas no corpus. Se as alternâncias definidas manualmente para a classe forem as mesmas encontradas no corpus para um determinado verbo candidato, este verbo passa a ser membro da VerbNet.Br; caso contrário o verbo é descartado. No

contexto da Linguística de Corpus, a segunda etapa é independente e de extrema importância. É independente, pois pode ser realizada isoladamente e produz resultados próprios que podem ser utilizados para outras tarefas. É de extrema importância, pois, sabendo a ocorrência das alternâncias sintáticas dos verbos em um corpus da língua portuguesa, será possível comparar com as traduções do inglês, identificando o que há de comum e o que há de diferente entre as línguas. Para apoiar a realização desta etapa serão estudados trabalhos com subcategorization frames como Zanette (2010) e Altamirano e Alemany (2010).

Contato: carol.scarton@gmail.com

CORPORA NA IDENTIFICAÇÃO DE COLOCAÇÕES VERBAIS PARA UM DICIONÁRIO BILÍNGUE

Danilo Murakami (IC/USP)

Segundo Tagnin (1999, p. 400), uma colocação é uma combinação lexical recorrente, não-idiomática, coesiva, arbitrária, cujos constituintes são contextualmente restritos. Uma colocação verbal, portanto, é a associação de um verbo a um adjetivo ou a um substantivo, o qual pode ter função de objeto ou sujeito. Nossa pesquisa está inserida no projeto de um dicionário bilíngue, inglês-português/português-inglês, de colocações verbais. A coleta de dados para o dicionário teve inicío em um momento prévio ao uso de corpora, tendo como fonte outros dicionários, material escrito, e até mesmo a competência linguística dos pesquisadores envolvidos na coleta (Tagnin, 2005, p. 201). O trabalho aqui apresentado limita-se à verificação do estatuto de colocação e à padronização dos termos coletados para figurarem entradas de verbete de dicionário, por meio da exploração de corpora, a fim de retratar a maneira pela qual a combinação entre verbo e substantivo/adjetivo ocorre, uma vez que diversas estruturas são possíveis (Ibid., p. 203-205). Atualmente o dicionário conta com mais de 4500 entradas nos dois idiomas, das quais uma média de 3200 já foram verificadas de modo a padronizar a entrada do verbete, eliminando por volta de 300 casos que não possuíam as características de uma colocação acima descritas. Cada dado coletado é submetido a uma pesquisa inicial no Google a fim de verificar se apresenta um número de ocorrências relevante. A comprovação dos termos como colocação também é feita por meio de corpora on-line. Para o inglês, temos utilizado o Corpus of Contemporary American English (Davies, 2008), e para o português, o Corpus do Português (Davies & Ferreira, 2006). Por serem morfologicamente etiquetados, ambos os corpora facilitam a identificação da estrutura da colocação. Dessa forma, por exemplo, a entrada "não ver com bons olhos" foi alterada para "ver com bons olhos", visto que ocorrências em que o verbo é usado afirmativamente também foram encontradas. Já "intentar ação judicial" foi substituído por "intentar uma ação", pois outros tipos de ações também são intentadas conforme mostram dados do corpus. Esperamos que a pesquisa por meio de corpora possa auxiliar a identificação das estruturas dos demais termos coletados, assim como sugerir outras colocações ainda não identificadas.

Contato: danilosuzuki@gmail.com

Referências:

DAVIES, M. (2008-) The Corpus of Contemporary American English (COCA): 425 million words, 1990-present. Disponível em http://www.americancorpus.org.

DAVIES, M. & FERREIRA, M. (2006-) Corpus do Português (45 million words, 1300s-1900s). Disponível em http://www.corpusdoportugues.org.

TAGNIN, S. E. O. Collecting data for a bilingual dictionary of verbal collocations: from scraps of paper to corpora research. In: PALC '99 - International Conference on Practical Applications in Language Corpora, 2000, Lodz. PALC '99: Practical Applicatons in Language Corpora. Papers from the International Conference at the University of Lodz, 15-18 April 1999. Lodz : Universidade de Lodz, 1999. p. 399-407.

TAGNIN, S. E. O. Um dicionário de colocações verbais? Para quê?. In: Tony Berber Sardinha. (Org.). A Língua Portuguesa no Computador. Campinas: Mercado de Letras Edições e Livraria Ltda., 2005, p. 197-214.

ENSINO DE INTERPRETAÇÃO SIMULTÂNEA NA GRADUAÇÃO: UMA ANÁLISE DE CORPORA DE APRENDIZES

Luciana Latarini Ginezi (USP)

O ensino de Interpretação em Universidades brasileiras tem crescido a cada ano. No entanto, pouco se sabe a respeito dos métodos utilizados por professores de interpretação, que normalmente reproduzem a maneira como aprenderam ou utilizam exemplos de sua prática na profissão. Este projeto pretende dar um passo em direção à pesquisa em ensino de Interpretação, discutindo se o ensino de Interpretação Simultânea vem sendo ensinado no momento oportuno, após a Consecutiva, ou se poderíamos inverter essa ordem. Utilizando a metodologia da Linguística de Corpus para investigar o produto de interpretações de alunos de alguns cursos de Tradução e Interpretação da cidade de São Paulo, demonstraremos as diferenças e semelhanças entre alunos iniciantes e concluintes, em relação aos padrões linguísticos e estratégias encontradas nos corpora de aprendizes. Faremos, ainda, a análise do processo desenvolvido, utilizando entrevistas retrospectivas, que posteriormente serão comparadas às análises de corpora para obtenção de resultados conclusivos finais.

Contato: luginezi@uol.com.br

AS ORAÇÕES EXISTENCIAIS NO PORTUGUÊS BRASILEIRO: UM ESTUDO EXPLORATÓRIO SOB A PERSPECTIVA DA LINGÜÍSTICA DE CORPUS E DA TEORIA SISTÊMICO-FUNCIONAL

Fabiane Santos (UFMG)

Desenvolvido no âmbito do grupo de pesquisa Modelagem sistêmico-funcional da tradução e da produção textual multilíngue, do Laboratório Experimental de Tradução (LETRA) da Faculdade de Letras/UFMG, este trabalho adota uma perspectiva sócio-semiótica para descrição e análise de

orações existenciais em português brasileiro, isto é, de orações que realizam gramaticalmente significados de existência e acontecimento (Halliday e Matthiessen, 2004). Mais especificamente, este trabalho parte da análise de diversos tipos de textos em que há ocorrência de orações existenciais, visando descrever variáveis que possam indicar padrões sistêmicos capazes de oferecer dados para a comparação, o contraste e a investigação da relação tradutória entre os processos existenciais em pares lingüísticos que contemplem o português brasileiro. Os dados são extraídos do corpus CALIBRA (Catálogo da Língua Brasileira), desenvolvido na FALE/UFMG, cuja compilação contempla o funcionamento da língua em seu contexto de cultura (Halliday, 1978). O corpus possui um milhão de palavras (tokens) distribuídas em textos agrupados em oito processos sócio-semióticos (Ure, 1969a e 1969b), a saber: explicar, reportar, recriar, compartilhar, fazer, recomendar, habilitar e explorar (Matthiessen et al., 2008). Com as ferramentas do software WordSmith Tools (Scott, 2007), foram extraídas linhas de concordância referentes a ocorrências de orações existenciais, as quais foram analisadas manualmente a partir dos pressupostos da teoria sistêmica (Halliday, 2002) e da descrição sistêmico-funcional do português feita por Figueredo (2011). Uma vez concluída a análise dos dados do corpus brasileiro, foram feitas buscas por orações existenciais no corpus Compara (Frankenberg-Garcia & Santos, 2000), na direção inglês-português, e as ocorrências achadas nos originais em inglês foram analisadas para observar de que forma as mesmas foram traduzidas para o português. Os resultados obtidos do corpus paralelo foram relacionados com os resultados obtidos do corpus do português brasileiro não traduzido, em particular com aqueles obtidos para o subprocesso recriar, com o objetivo de verificar se os significados existenciais construídos nos textos em português traduzidos guardam analogia com aqueles observados no corpus de português não traduzido.

Contato: fabianechuck@gmail.com

SIGNIFICADOS EXISTENCIAIS NO PORTUGUÊS BRASILEIRO: UM ESTUDO CONTRASTIVO EM TEXTOS TRADUZIDOS E NÃO TRADUZIDOS

Kícila Ferreguetti (IC/UFMG)

Este trabalho está inserido no grupo de pesquisa Modelagem sistêmico-funcional da tradução e da produção textual multilíngue, desenvolvido no Laboratório Experimental de Tradução (LETRA) da Faculdade de Letras/UFMG, e tem como objetivo contribuir, a partir da perspectiva sócio-semiótica, para a descrição e análise de orações existenciais no português brasileiro. Segundo Halliday e Matthiessen (2004), os processos existenciais, ainda que não possuam uma frequência de ocorrência alta no discurso (se comparados aos demais processos), são relevantes na construção dos diferentes tipos de texto, uma vez que realizam gramaticalmente a existência e o acontecimento. Este trabalho visa analisar as ocorrências de orações existenciais em diversos tipos de textos, com o intuito de identificar variáveis e padrões sistêmicos que possam ser generalizados e utilizados como ferramenta comparativa no campo dos estudos da tradução. Para isso, partindo-se da premissa de que a língua funciona dentro de um contexto (Halliday, 1978) e que seu estudo pode ser abordado com base nos seus processos sócio-semióticos, a saber: compartilhar, explicar, explorar, fazer, habilitar, recomendar, recriar e reportar (Ure, 1969a e 1969b) (cf. Matthiessen et al., 2008), foram feitas buscas de orações que realizam significados existenciais no corpus CALIBRA

(Catálogo da Língua Brasileira), desenvolvido na FALE/UFMG. O corpus é composto por textos representativos dos oito processos sócio-semióticos acima mencionados, cada um contendo 125.000 palavras (tokens), totalizando um milhão de tokens. Com o auxílio do software WordSmith Tools (Scott, 2007), foram geradas linhas de concordâncias envolvendo realizações de significados existenciais, as quais foram analisadas manualmente, segundo os pressupostos da teoria sistêmico-funcional (Halliday, 2002) e a descrição sistêmico-funcional do português feita por Figueredo (2011). Após a conclusão desta primeira parte da análise, foi selecionado um corpus paralelo português-italiano, composto de originais e traduções de obras da autora Clarice Lispector, integrantes do corpus CORDIALL (FALE/UFMG), com o objetivo de identificar e verificar como as orações existenciais presentes nos originais foram traduzidas para o italiano.

Contato: kicilaferreguetti@yahoo.com.br

OS QUANTIFICADORES A FEW E (VERY)FEW: QUESTÕES DE INTERLÍNGUA E PROSÓDIA SEMÂNTICA EM CORPUS DE APRENDIZES

Rejane Protzner Silero (IC/UFMG)

Este trabalho se enquadra na proposta de Bennet (2010) no tocante à contribuição oferecida pela observação de linhas de concordância para responder questões diversas inseridas na dimensão da fraseologia (colocações, agrupamentos lexicais, etc.) No caso específico desse estudo, realizou-se algo aos moldes do que foi produzido por Ruzaité (2009), cujo trabalho mostrou características do sistema de quantificadores em inglês e em lituano, utilizando um corpus paralelo como base. No entanto, o presente trabalho difere do de Ruzaité por ter seus dados gerados a partir de um corpus de aprendizes, que propulsou uma análise de inadequações linguísticas envolvendo o uso de a few e (very)few em produções textuais de alunos brasileiros de graduação em Letras na UFMG, habilitação Inglês. Assim sendo, foram elaboradas uma hipótese linguística e uma hipótese pedagógica. A primeira é que a prosódia semântica (Sinclair, 1987) mais restritiva e negativa no uso de (very)few não é capturada pelos aprendizes, que acabam utilizando quantificadores distintos de forma intercambiável. Já a segunda recupera a dicotomia descrição vs. prescrição, propondo que os aprendizes devem ser mais claramente conduzidos às diferenças existentes entre o uso dos quantificadores em português brasileiro e em inglês. Além disso, é esperado que, por meio da aplicação de atividades baseadas em linhas de concordância, eles se tornem usuários mais bem-sucedidos da língua-alvo, atentando-se para padrões léxico-gramaticais (Conrad, 2000). A metodologia utilizada resumiu-se no uso da seção acadêmica da plataforma de Davies (versão da língua portuguesa) e, como corpora de referência, o LOCNESS - compilado por Granger - e a seção acadêmica do corpus geral COCA foram usados. A conclusão preliminar é a de que o ensino pautado em uma conduta léxico-gramatical pode auxiliar no tratamento a questões de interlíngua. Dessa forma, a utilização de ferramentas da LC se torna de grande importância dentro de sala.

Contato: rejaneprotzner@gmail.com

MAPEAMENTO DAS ORAÇÕES EXISTENCIAIS NO PORTUGUÊS BRASILEIRO

Adriana Pagano (UFMG)
Giacomo Figueredo (UFMG)

Na lingüística sistêmico-funcional (Halliday & Matthiessen, 1999, 2004), as orações existenciais constroem significados relativos ao surgimento ou aparecimento de seres ou entidades no discurso e ao acontecimento de processos e eventos. Uma das suas principais funções é a de introduzir um novo referente no discurso, o qual será retomado e sobre o qual outros significados serão construídos, uma vez que o mesmo passa a existir. Para Halliday & Matthiessen (2004), as orações existenciais cumprem um papel relevante em determinados tipos de texto, como é o caso das histórias, nas quais introduzem participantes no movimento inicial da narrativa, ou em guias turísticos, nos quais apresentam lugares e pontos de interesse sugeridos para visitação do público leitor. Além desses dois tipos de texto, presume-se que os significados existenciais desempenhem papeis relevantes em outros tipos textuais, embora estudos a respeito sejam praticamente inexistente na literatura. No que diz respeito ao repertório textual dos usuários da língua portuguesa do Brasil, pouco se conhece sobre a ocorrência e características das orações existenciais em diferentes tipos textuais. Este trabalho visa preencher essa lacuna através de um estudo exploratório das orações existenciais no português brasileiro, utilizando- se subsídios da lingüística de corpus. Os dados são gerados através de buscas no corpus CALIBRA (Catálogo da Língua Brasileira), desenvolvido na FALE/UFMG, composto por textos representativos de oito processos sócio-semióticos, cada um representado no corpus por textos que perfazem 125.000 palavras (tokens), totalizando um milhão de tokens. O objetivo é mapear a freqüência de ocorrência de orações existenciais por processo sócio-semiótico e analisar seu papel no desenvolvimento dos textos vinculados a cada processo. O trabalho está inserido no grupo de pesquisa Modelagem sistêmico-funcional da tradução e da produção textual multilíngue, desenvolvido no Laboratório Experimental de Tradução (LETRA) da Faculdade de Letras/UFMG.

Contato: apagano@ufmg.br

CONSTRUÇÕES COM TO E FOR EM PRODUÇÕES ESCRITAS DE APRENDIZES DA LÍNGUA INGLESA - O USO DAS FERRAMENTAS DA LINGUÍSTICA DE CORPUS

Marlei Rose Renzetti Tartoni (Poslin/UFMG)

As produções de texto de aprendizes da língua Inglesa estão sendo, cada vez mais, objetos de estudo e análise dentro de diversas áreas da linguística aplicada (Granger, 1998, Bernardini, 2004). A compilação de corpora de diversos tamanhos e o uso de softwares específicos dentro da Linguística de Corpus tornaram possível o acesso facilitado aos dados extraídos das produções de aprendizes, organizados em listas de palavras, linhas de concordância, clusters e outras disponibilizações que giram em torno de itens lexicais que se transformam em objeto de pesquisa e análise, dentro de processos de investigação e/ou confirmação de hipóteses (Johns, 1994, Conrad, 2000). Produções lingüísticas são, portanto, solicitadas em diversas instituições de ensino

e de formas variadas. Esta pesquisa, objeto de dissertação de tese de mestrado de Marlei Rose Renzetti Tartoni, iniciado em fevereiro de 2010, com previsão para defesa em fevereiro de 2012, foi orientada por Deise Prina Dutra, professora doutora da Universidade Federal de Minas Gerais e encontra-se em andamento. Teve como objetivo principal a avaliação do aumento de proficiência de dois grupos de alunos do 9º ano do ensino fundamental de uma escola estadual de Minas Gerais, submetidos a exercícios com construções lingüísticas formadas com os itens to e for. Os participantes têm entre 13 e 16 anos, divididos em 17 do sexo feminino e 19 do sexo masculino. Ambos os grupos escreveram piadas e biografias e participaram de atividades de leitura com textos desses gêneros e, posteriormente, fizeram exercícios linguísticos diferentes. O grupo controle foi submetido a exercícios tradicionais de gramática com preenchimento de lacunas com os itens to e for e o grupo de tratamento fez atividades de conscientização linguística baseadas em linhas de concordância de textos dos gêneros piada e biografia. Essas linhas de concordância foram retiradas de um pequeno corpus compilado pela pesquisadora que contem 45 piadas e 55 biografias, sendo que há 4958 palavras no primeiro subcorpus e 19614 no segundo. Os dois grupos produziram textos, piadas e biografias, antes e após os exercícios aplicados que foram analisados como pré e pós-teste. As produções foram digitalizadas e analisadas com o programa WordSmith Tools 5.0. Os resultados desta pesquisa pretendem demonstrar como os dois grupos utilizaram as construções linguísticas com to e for, com diferentes funções pragmáticas antes e após as atividades propostas de gramática tradicional e baseadas em corpora. A análise preliminar aponta que o grupo tratamento fez um maior uso das construções em questão, melhor estruturadas e com maior variedade de funções pragmáticas.

Contato: mrrtartoni@gmail.com

Referências:

BERNARDINI, Silvia. Corpora in the Classroom: An Overview and Some Reflections on Future Developments. In: SINCLAIR, John McH. (ed.). How to Use Corpora in Language Teaching. Philadelphia: John Benjamins Publishing Company. 2004. p.15-36.

CONRAD, Susan. Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century? TESOL Quarterly, v. 34. n. 3, p. 548-560, 2000.

GRANGER, Sylviane. TRIBBLE, Chris. Learner Corpus in the Foreign Language Classroom: Form-Focused Instruction and Data-Driven Learning. In: GRANGER, S. (Ed.) Learner Corpora on Computer. London: Longman. 1998. p. 199-209.

JOHNS, Tim. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In: ODLIN, R. Perspectives on Pedagogical Grammar. Cambridge: Cambridge. 1994. p. 293-313.

TEMA PREDICADO NO PORTUGUÊS BRASILEIRO: CONTRIBUIÇÕES DE UMA ABORDAGEM DE CORPUS PARA SEU ESTUDO

Giacomo Figueredo (UFMG)
Adriana Pagano (UFMG)

Na lingüística sistêmico-funcional (Halliday & Matthiessen, 2004), o Tema predicado (predicated Theme) é caracterizado como um recurso léxico-gramatical que introduz um contraste nos significados construídos no texto. Uma de suas funções é dar destaque a adjuntos circunstanciais de modo, de localização espacial e temporal, entre outros, os quais são tematizados em pontos específicos, de forma a mudar o fluxo discursivo e re-direcioná-lo de acordo com os significados aos quais se deseja dar maior intensidade (Matthiessen, 1995). São escassos os estudos do Tema predicado sob uma perspectiva sistêmico-funcional, sobretudo no que diz respeito a seu uso na língua portuguesa do Brasil. Diante dessa lacuna, este trabalho se propõe a explorar o uso do Tema predicado a partir da observação de ocorrências levantadas de um corpus de textos representativos de processos sócio-semióticos da língua portuguesa do Brasil. O corpus CALIBRA (Catálogo da Língua Brasileira), desenvolvido na FALE/UFMG, possui aproximadamente um milhão de tokens e contempla textos orais e escritos pertencentes a diferente tipos textuais e registros. A análise das ocorrências extraídas do corpus é feita dentro do âmbito da descrição sistêmico-funcional iniciada por Figueredo (2011), mais especificamente em sua relação com o sistema de Tema em português brasileiro, elaborado pelo autor. O trabalho está inserido no grupo de pesquisa Modelagem sistêmico-funcional da tradução e da produção textual multilíngue, desenvolvido no Laboratório Experimental de Tradução (LETRA) da Faculdade de Letras/UFMG.

Contato: giacomojakob@yahoo.ca

PACOTES LEXICAIS: ASPECTOS LINGUÍSTICOS DA INTERLÍNGUA EM CORPUS DE APRENDIZES DE LE DO ENSINO MÉDIO

Shirlene Bemfica De Oliveira (UFMG)
Kamila Oliveira Carmo (UFMG)
Tatiane Morandi de Oliveira (UFMG)
Amanda Rossi (UFMG)
Ivan Inacio (UFMG)

Este trabalho tem como objetivos mapear e descrever os pacotes lexicais (lexical bundles) típicos de alunos iniciantes evidenciados em corpus escritos de textos argumentativos. O estudo de caso foi desenvolvido com a participação da pesquisadora, quatro alunos bolsistas do ensino médio e aproximadamente 230 alunos da segunda série distribuídos em sete turmas do Ensino Médio de um Instituto Federal em Minas Gerais. Os dados estão sendo coletados por meio de uma produção de um argumentativo escrito pelos alunos e as análises feitas com o auxílio da ferramenta AntConc e da categorização proposta por Simpson-Vlach e Ellis (2010). A análise quantitativa será feita com base na freqüência e função dos itens investigados. “As técnicas quantitativas são

essenciais para a pesquisa baseada em corpus", pois elas auxiliam na compreensão do comportamento das palavras em determinados contextos de uso e freqüência, além de respaldar e enriquecer as análises (BIBER, 1998, p. 8). Este tipo de pesquisa possibilita chegar à linguagem produzida pelos alunos por meio da análise dos padrões probabilísticos que se constroem nos contextos de uso (SARDINHA, 2000). Por meio desta abordagem é possível mapear as características do discurso típico de aprendizes iniciantes e a investigação das freqüências dos traços lingüísticos (fórmulas e grupos lexicais), pois a comprovação da freqüência atestada é que levará o pesquisador a probabilidade teórica (SARDINHA, 2004). Os resultados iniciais mostram que mesmo em níveis iniciais, os alunos utilizam fórmulas e pacotes lexicais no uso da língua e apontam para a relevância do ensino dos pacotes lexicais no contexto escolar.

Contato: shirleneo@yahoo.com

ANÁLISE DE ORAÇÕES RELATIVAS EM CORPUS DE APRENDIZES DO ENSINO MÉDIO

Shirlene Bemfica De Oliveira (UFMG)
Ana Rachel Simões Fortes (UFMG)
Maria Teresa de Andrade Sol (UFMG)
Gabriela Leite (UFMG)
Pamela Felix (UFMG)

Este projeto tem como objetivo analisar linhas de concordâncias em corpus de aprendizes iniciantes e mapear as orações relativas em língua inglesa. Adotamos o conceito de orações relativas de Biber et. al. (1999) e os princípios da instrução formal como foco na forma proativo (Ellis, 1994, 2001, Doughty e Williams, 1998, Pieneman, 1998, Doughty, 2001, Schimidt, 2001 e Williams, 2001). Através desta investigação analisaremos a influência de uma abordagem com foco na forma como instrumento para aumentar a incidência de noticing dos alunos através de tarefas que selecionam a atenção. O objetivo e compreender o processo de aquisição das orações relativas em LI pela investigação dos efeitos de uma intervenção pedagógica com foco na forma no que diz respeito ao uso das orações relativas em língua inglesa. O estudo de caso, de natureza qualitativa, está sendo desenvolvido contando com a participação as pesquisadoras, quatro alunos bolsistas do ensino médio e sete turmas do Ensino Médio. Os dados foram coletados por meio de tarefas de produção escrita de textos expositivos de definição. A análise dos dados feita com o auxílio do programas AntConc que permite a identificação de padroes linguisticos com as ocorrências das orações relativas, a análise da composição lexical, a temática dos textos selecionados e a organização retórica e composicional do gênero discursivo (SARDINHA, 2004). Os resultados apontam para o aumento da freqüência dos pronomes relativos e para o aumento da compreensão das relativas.

Contato: shirleneo@yahoo.com

X Encontro de Linguística de Corpus
V Escola Brasileira de Linguística Computacional
9 a 12 de Novembro de 2011
Faculdade de Letras / UFMG

A Chavicidade Na Análise De Estilo Em Tradução: Um Estudo Baseado Em Corpora Paralelos Espanhol/Português

Ariel Novodvorski (PosLin-UFMG/UFU)
Célia Magalhães (UFMG)

As pesquisas sobre estilo em tradução vêm definindo um campo de estudos para a reflexão e análise dos mais diversos aspectos linguísticos nos textos traduzidos. A partir de trabalhos pioneiros, apoiados nos subsídios advindos da pesquisa baseada em corpus, estudiosas como Baker (2000), Malmkjaer (2003; 2004; 2005) e Bosseaux (2004; 2007), entre outros, vêm demonstrando interesse por questões específicas de estilo em tradução e por discussões sobre a (in)visibilidade do tradutor. Graças aos recursos que provêem as ferramentas da linguística de corpus, as pesquisas foram se especializando na investigação da presença discursiva do tradutor, tradição que se foi edificando a partir dos trabalhos de May (1994), Venuti (1995), Hermans (1996) e Schiavi (1996). Também Munday (2008) está inserido nesse contexto de pesquisa, em função do estudo de aspectos de ideologia vinculados ao estilo em tradução, especificamente na escrita latino-americana traduzida à língua inglesa, com os subsídios da análise de registro, na perspectiva hallidayana. No âmbito nacional, Magalhães (2005), Camargo (2009) e recentemente Barcellos e Magalhães (2011) investigam padrões de estilo na tradução literária. Este trabalho está afiliado a esse marco de investigação, constituindo uma pesquisa de doutorado em andamento sobre as variações estilísticas na tradução literária brasileira. O corpus ESTRA - Estilo em tradução, desenvolvido no âmbito do LETRA/FALE/UFMG, é o material de análise, especificamente um subcorpus paralelo bilíngue, composto por três traduções feitas pelo tradutor Sérgio Molina, da língua espanhola ao português brasileiro, de três obras do autor argentino Ernesto Sabato. No escopo deste trabalho, concentra-se a atenção nas variações observadas a partir do levantamento e análise das palavras-chave, com especial atenção às agrupações lexicais (clusters) e colocados de alguns dos termos pesquisados, em função de sua chavicidade, e levando em consideração as mudanças tradutórias observadas. Apresentam-se brevemente os procedimentos metodológicos para a compilação, preparação e etiquetamento do corpus de análise. Resultados parciais apontam para a presença da voz do tradutor e para possíveis marcas de seu estilo, observadas a partir da aplicação das ferramentas básicas empregadas nos estudos baseados em corpus. Observa-se uma linguagem mais variada de tradução e uma tendência à explicitação de determinados elementos, nos textos traduzidos.

Contato: arielnovodvorski@yahoo.com.br