ISSN 1517-9257
Campo Grande | MS
Vol. 18 | Nº 35 | 2014

Papéis
REVISTA DO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS DE LINGUAGENS - UFMS

Colocações especializadas na ciência da computação: uma proposta de estudo terminológico para fins de ensino de inglês

Specialized collocations in computer science: a terminological study proposal for english teaching

Andrea Monzón[*]
Ana Paula Lemke[**]
Juliane de S. N. de Moura[***]

Resumo: O objetivo deste trabalho é propor um estudo descritivo das colocações de cunho terminológico em artigos científicos em inglês no campo da Ciência da Computação. Assim, a razão do desenvolvimento deste projeto ocorre pela necessidade de estudantes e professores não nativos de reconhecer as convencionalidades léxicas ao lerem em tal língua-alvo. Essa demanda se constitui no âmbito da educação profissional técnica e tecnológica em Informática, na qual a leitura de textos especializados anglófonos é parte integrante da formação dos alunos. Esses aprendizes têm que não somente compreender um idioma, que é a *lingua franca* da ciência, como também apreender conceitos e terminologias relevantes à prática de seu ofício. À luz da Linguística de *Corpus*, bem como de teorias da Terminologia e dos estudos do léxico, foram levantadas as colocações especializadas através de duas ferramentas computacionais. A partir dos dados coletados, esses padrões lexicais foram observados em seus contextos de uso. Conclui-se ser relevante dar continuidade às análises realizadas neste estudo preliminar, uma vez que as mesmas demonstraram ser produtivas para o ensino de inglês instrumental no panorama educacional mencionado.

Palavras-chave: colocações especializadas. artigo científico. terminologia. linguagens especializadas.

[*] Universidade Federal do Rio Grande do Sul/IFRS, andrea.monzon@feliz.ifrs.edu.br
[**] Instituto Federal do Rio Grande do Sul, ana.lemke@feliz.ifrs.edu.br
[***] Instituto Federal do Rio Grande do Sul, juliane.moura@feliz.ifrs.edu.br

Abstract: *The objective of this paper is proposing a descriptive study of terminological collocations in Computer Science research articles in English. The reason to develop it occurs due to non-native students' and teachers' necessity to recognize the lexical conventionalities when reading in such target language. This need happens in the situation of technical and technological professional education of Informatics, in which reading Anglophone specialized texts is part of students' training. These learners have not only to understand a language, that is the "lingua franca" of science, but also figure out relevant concepts and terminologies to their job practice. Based on Corpus Linguistics as well as Terminology theories and lexicon studies, specialized collocations were obtained using two computational tools. From the data raised, these lexical patterns were observed in their contexts of use. In conclusion, it is relevant to continue the analyses made in this preliminary study, once they demonstrated to be productive for English for Specific Purposes teaching in the educational context mentioned.*

Keywords: *specialized collocations. research article. terminology. specialized languages.*

Introdução

No âmbito da comunicação internacional acadêmica e científica, o inglês é a *lingua franca*. É através de tal idioma que pesquisas são divulgadas e compartilhadas com estudantes, professores e profissionais de todo o mundo. O meio mais utilizado para comunicar as realizações de um estudo é o artigo científico, o qual, como maneira de inserção na comunidade acadêmica de forma mais efetiva, utiliza-se da Língua Inglesa (doravante LI).

Para atender a uma comunidade específica e especializada, o artigo científico pressupõe que sua forma e seu conteúdo atendam a convencionalidades linguísticas. Essas não são tão triviais, principalmente para leitores iniciantes, que além de não ter domínio do gênero textual em si, desconhecem como as palavras se combinam apropriadamente.

"Falante ingênuo", conceito de Fillmore (1979 apud ALVES; TAGNIN, 2011), em muito auxilia o entendimento de como o aprendiz principiante compreende uma língua estrangeira de forma literal. Isso faz com que ele faça escolhas lexicais equivocadas ao falar e escrever e estabeleça uma correlação unívoca das palavras de sua língua materna com a estrangeira. Esse desconhecimento das convenções linguísticas torna evidente a não proficiência e a demanda por recursos pedagógicos - dicionários gerais, glossários e ferramentas computacionais - que auxiliem de maneira efetiva os usuários não nativos. No meio acadêmico, as dúvidas dos não nativos podem ser ainda

maiores, pois suscitam o uso de léxico especializado dentro de gêneros textuais sofisticados e que remetem a uma comunidade discursiva exigente.

No presente trabalho, para fins de um estudo piloto, faz-se o levantamento e a análise de ocorrências do que aqui se denomina colocações especializadas, bem como seus usos em textos da área de Ciência da Computação, dentro do gênero artigo científico (SWALES, 1990). Para tanto, foi utilizado aporte teórico da Linguística de *Corpus* e das teorias da Terminologia, além de subsídios metodológicos do Processamento de Língua Natural (PLN). O intuito é verificar as possíveis contribuições desse tipo de recurso lexical para o ensino de inglês instrumental (HUTCHINSON, WATERS, 1987) no âmbito da educação profissional técnica e tecnológica em Informática.

Seja em forma de bases de dados, repositórios, glossários, dicionários ou ferramentas pedagógicas, a arbitrariedade das CEs deve ser contemplada em produtos terminológicos e/ou terminográficos que atendam a usuários reais com as dificuldades já discutidas. Essa demanda ocorre fortemente no ensino de inglês instrumental (HUTCHINSON; WATERS, 1987) em cursos técnicos e tecnológicos da Educação Profissional. Nessa situação, além de aprender a LI e suas peculiaridades morfológicas e sintáticas, os alunos devem realizar leitura de textos especializados e/ou artigos científicos, a fim de construir sua formação e constante atualização para um ofício porvir. Esses textos contêm CEs que remetem a conceitos que ainda estão sendo aprendidos pelos alunos nas disciplinas técnicas.

No âmbito do ensino técnico e tecnológico do Brasil, especialmente nos cursos de nível médio, existe uma grande demanda de produção de materiais pedagógicos que possam auxiliar alunos e professores na leitura de artigos científicos em inglês, sobretudo no cenário dos Institutos Federais. A área em que há o maior número de alunos matriculados na educação profissional de nível médio é a Informática, chegando a cerca de 10% (INEP, 2013a). Já, no nível de Ensino Superior, os cursos tecnológicos em Ciência da Computação e afins despontam entre os cinco mais procurados há alguns anos (INEP, 2013b). É esse o panorama para o qual esta proposta se dispõe a contribuir.

A esta introdução, seguem-se quatro partes. Na seção 2, são apresentados os pressupostos teóricos que fundamentam este trabalho. Na seção 3, são explanados os métodos empregados, no que diz respeito ao *corpus* estudado e às ferramentas computacionais utilizadas. Na seção 4, os resultados são elencados e discutidos, demonstrando como ocorrem as colocações especializadas em artigos científicos da área de Ciência da Computação. Já na seção 5, estão dispostos alguns encaminhamentos possíveis quanto ao porvir deste estudo.

2 Contribuições da Linguística de *Corpus*

Devido ao constante aumento da capacidade de armazenamento e processamento dos computadores, nas últimas décadas a Linguística de *Corpus* (doravante LC) cada vez mais tem sido difundida e empregada em diferentes áreas, seja como abordagem ou metodologia (FINATTO; EVERS; ALLE, 2010). Isso não é uma novidade do século XX, pois já na Antiguidade e Idade Média havia corpora de trechos da Bíblia, muito embora ainda não existisse o que hoje se conhece por computador, nem houvesse pressupostos teóricos muito definidos. O que existia de semelhante entre esses estudos e aqueles desenvolvidos por Thordike, West e Quirk no século XX, era uma preocupação em conhecer melhor as línguas observando como e com que frequência as palavras eram utilizadas (SARDINHA, 2000). Durante séculos, antes de se estruturar de maneira teórico-metodológica, a LC produziu listas de vocábulos, a fim de documentar e quantificar o patrimônio lexical de algumas comunidades linguísticas, principalmente, no que dizia respeito ao inglês. O que se tem, atualmente, é um panorama bem mais vasto de atuação em pesquisa e de produção de recursos para usuários reais.

Logo, a LC não está somente interessada em quantificar palavras, pois também há a preocupação em observar e compreender seus usos. Assim, a língua é entendida como um sistema probabilístico (HALLIDAY; ANGUS; STREVENS, 1965), buscando-se identificar padrões que a descrevam como ela ocorre no mundo *in vivo* (OLIVEIRA, 2009). Essa não é uma nova

linguística, mas um novo caminho para a Linguística (FINATTO; EVERS; ALLE, 2010).

Neste estudo, *corpus* é entendido como um conjunto de documentos em formato eletrônico construído para um propósito específico (ALUÍSIO; ALMEIDA, 2006; SARDINHA, 2000), assim como o sintetizam McNery e Wilson (1997, p.24):

> Então um corpus na Linguística Moderna, em contraste a simplesmente ser qualquer corpo de texto, deve ser mais apropriadamente descrito como um conjunto de texto de tamanho finito e computacionalmente legível, compilado de maneira a representar maximamente a variedade linguística levada em consideração. Entretanto, o leitor deveria estar atento às possibilidades de desvios, em certas situações, dessa definição 'prototípica'.

O planejamento do *corpus* é uma fase de suma importância em qualquer estudo que tenha tal conjunto como fonte inicial e/ou principal de seus dados. Critérios como autoria, conteúdo, modo, tempo e finalidade delineiam os caminhos a serem percorridos pelo pesquisador (SARDINHA, 2000; 2004). Caso essas características não sejam profundamente planejadas, o levantamento e a análise das ocorrências podem ser prejudicados. Outros aspectos, ainda, são a representatividade, extensão, especificidade e adequação do *corpus*.

O *corpus* aqui estudado será descrito na Seção 3, bem como os critérios estabelecidos para a sua compilação, a fim de atender os propósitos desta pesquisa.

2.2 Gênero artigo científico e PLN

A dificuldade de não nativos em redigir artigos científicos em inglês é uma demanda com a qual várias áreas do conhecimento se preocupam. A fim de divulgar suas pesquisas e aprimorar seus currículos, estudantes e professores escrevem artigos em português contendo um *abstract* ou inteiramente em inglês. Essa tarefa não é fácil, uma vez que em tal empreitada, brasileiros são considerados não nativos, ou seja, constantes aprendizes. Além das convenções que dizem respeito à sua estrutura esquemática bem como

suas características estilísticas e de apresentação, o artigo científico (SWALES, 1990) possui um uso mais sofisticado e formal das palavras, uma vez que se insere em comunidades discursivas especializadas.

Sendo o gênero originário da atividade humana, sua riqueza e variedade são infinitas. Cada um deles tem no mundo uma função mais ou menos sofisticada, por isso Bakhtin (1997) os classificou em gêneros primários e secundários. Os denominados de primários são gêneros mais simples, estabelecendo uma relação direta com a realidade em que se inserem. Os secundários ocorrem “em circunstâncias de uma comunicação cultural mais complexa e relativamente mais evoluída, principalmente escrita” (BAKHTIN, 1997, p.281). Como todo texto tem um sujeito, esse busca, através de suas escolhas dentro das especificidades de um gênero, dar individualidade a seus enunciados, pois se o gênero tem uma função social, nada nele é dito ao acaso.

O texto concretiza um conjunto de necessidades de seu autor: motivação, finalidade e realização (KOCH, 2005). Os fatores sociais envolvidos na elaboração de um texto remetem aos seus conteúdos, à sua forma e sua intencionalidade. O artigo científico é um gênero textual e não um tipo textual (MARCUSCHI, 2002), possuindo um sistema interno que busca atender à sua proposta comunicativa (SWALES, 1990) de demonstrar aos leitores que: a) o(s) autor(es) conhece(m) as regras do jogo, ou seja, as convenções da forma e do conteúdo desse gênero; b) o(s) autor(es) tem(êm) domínio da área em que sua pesquisa está contida. Em contraposição, os tipos textuais têm uma natureza linguística que estabelece sua composição, como uma sequência de etapas teoricamente pré-estabelecidas, abrangendo as categorias narração, argumentação, exposição, descrição e injunção.

Não é trivial para um pesquisador redigir em inglês com tantas responsabilidades envolvidas. Buscando atender a essas demandas, algumas pesquisas foram desenvolvidas, no Brasil, com o intuito de atender às dificuldades reais e cotidianas de usuários não nativos de inglês no âmbito acadêmico, através de ferramentas computacionais. Em sua tese de

doutorado, Sandra Aluísio (1995) desenvolveu o AMADEUS (*Resources and Tools to help non-native English users in writing papers*), um conjunto de ferramentas de auxílio à escrita científica em inglês, alimentada com um *corpus* da área de Farmácia, e que possui módulos de suporte e crítica. Dando continuidade a esse projeto, foi concebido o CALeSE (*Computer-Aided Learning Tool Web-based for Scientific Writing in English*)[1], que possui um *corpus* categorizado e anotado, que inclui *abstracts* e introduções de textos de Ciência da Computação, Física e Linguística considerados bem escritos, e que foi desenvolvido pelo NILC (Núcleo Interinstitucional de Linguística Computacional), envolvendo duas universidades públicas paulistas.

Tais realizações se inserem em uma subárea da Inteligência Artificial denominada Processamento de Língua Natural (PLN), a qual surgiu devido à necessidade de tradução automática (*machine translation - MT*) em decorrência da II Guerra Mundial (NUNES, 2008). Nos anos 50, ocorreram os primeiros estudos institucionalizados, de Warren Weaver[2], que entendiam a tradução como algo como a criptografia da atualidade, ou seja, objetivava-se decifrar códigos (NUNES *et al*., 1999). Hoje se vê tal tratamento linguístico como inadequado, mas era o que se podia fazer em tal década, do ponto de vista computacional. Depois dessas tentativas iniciais, a MT passou a ser estudada de forma mais séria, tendo instituições de renome interessadas em viabilizá-la, como o Instituto de Tecnologia de Massachusetts (MIT) e as Universidades de Harvard, Califórnia e Georgetown. Esse foi só o início, pois o PLN não se resume a tradutores automáticos, uma vez que também se dedica ao desenvolvimento de outros tipos de ferramentas, tais como: *parsers*, sumarizadores automáticos, sentenciadores, tokenizadores, etiquetadores morfossintáticos, *chunkers* (GENOVÊS, 2007) e, mais recentemente, até mesmo simplificadores textuais (Projeto PorSimples[3] - Simplificação Textual do Português para Inclusão e Acessibilidade Digital), que têm a função social de incluir usuários através da leitura.

[1] http://www.nilc.icmc.usp.br/calese/
[2] Na época referida, Weaver era vice-presidente da Fundação Rockfeller.
[3] http://143.107.232.31/porsimples/index.php/P%C3%A1gina_principal

Esses recursos computacionais possibilitam a análise de grandes quantidades de textos tanto de forma quantitativa (*wordlist*, frequência, *keyword list*, lista de colocados) quanto qualitativa (*parsers*, concordanciadores, extratores de colocações). Através das mesmas, diversos estudos relevantes têm sido realizados em universidades brasileiras em trabalhos colaborativos envolvendo cientistas computacionais e linguistas. Essa parceria é bastante produtiva para mapear, descrever e compreender o uso das línguas. Claramente há, ainda, muito a ser feito em PLN, pois sendo uma área do conhecimento relativamente nova, está conquistando seu espaço aos poucos e tem mostrado sua importância em várias áreas do conhecimento.

2.3 Terminologia e texto especializado

A curiosidade e necessidade humanas de estudar seu léxico é algo que já existia desde os sumérios (2200 AC). Precisava-se nomear coisas no mundo para que as pessoas pudessem se comunicar. Com o avanço frenético da ciência no século XX, a demanda por nomear coisas que estavam sendo descobertas, inventadas e pesquisadas passou a ser uma preocupação, pois as nomenclaturas deveriam fazer com que cientistas de um mesmo país ou de países diferentes pudessem se entender e intercambiar conhecimento. Para atender a tais usuários das línguas, surgiu a Terminologia, que foi inaugurada pelo engenheiro elétrico Eugen Wüster, na Áustria dos anos 30 (KRIEGER; FINATTO, 2004). Ele entendia que cada palavra de uso especializado - termo - tinha com seu conceito uma relação unívoca, ignorando aspectos semânticos e pragmáticos, bem como a variação linguística. Embora seja reconhecível a importância de Wüster para dar início a um novo campo de estudo, sua visão de língua, principalmente do signo, transgride preceitos básicos da Linguística. Por essa e outras razões teórico-metodológicas, depois das contribuições iniciais desse vienense, estabelecendo a Teoria Geral da Terminologia (TGT), outras teorias se configuraram, construindo novos horizontes para a pesquisa terminológica.

Valorizando aspectos comunicativo-sociais, observando o conteúdo de um termo como algo variável, e estudando a língua *in vivo* e não *in vitro*, Cabré (2002) propôs a Teoria Comunicativa da Terminologia (ALMEIDA, 2006). Entretanto, ela optou por nomear seu objeto de estudo de "unidade terminológica" (UT). A pesquisadora espanhola se contrapõe à TGT, porque defende a implantação social dos termos em detrimento de sua normalização. Assim, a língua é necessariamente real, espontânea e natural, sem que se vise a padronizações. Não há, portanto, diferenciação entre palavra e termo, pois o que as distingue é sua situação comunicativa. Logo, o presente estudo faz-se valer do aporte teórico de Cabré, por corroborar com a perspectiva de que as UTs devem ser observadas dentro de seu ambiente natural, ou seja, nos discursos especializados, sendo neste caso o artigo científico.

Quando se pensa o texto como local de integração de componentes envolvendo a textualidade em si e a discursividade, o estudo aqui proposto se encontra com a Terminologia Textual. Essa observa o termo como elemento crucial dos textos especializados, sendo que esses são entendidos como aqueles que têm "termos e outros tantos elementos, igualmente passíveis de atenção" (FINATTO, 2008, p. 167). A investigação ocorre extrapolando os limites dos termos, centrando-se na descrição macro e microestrutural de *corpus* com apoio estatístico. Ciapuscio (1998) conduziu um estudo prototípico à luz dessa teoria, com o intuito de verificar os diferentes níveis de complexidade ocasionada pelo léxico em textos de 3 gêneros: artigo científico, artigo de divulgação científica e artigo de jornal. Cada um desses gêneros tem uma comunidade discursiva distinta e com um domínio mais ou menos sofisticado das temáticas abordadas. Na verdade, a autora teve a intenção de investigar a variação conceitual em Terminologia, e sua hipótese era de que "o grau de densidade da rede conceitual permite realizar afirmações fundamentadas acerca do nível de especialidade do texto"[4] (CIAPUSCIO, 1998, p. 1). Ela dividiu, assim, os leitores dos textos analisados em especialistas, semiespecialistas e leigos. O texto, para ela, constrói-se na interação social de acordo com o propósito comunicativo de seu emissor

[4] Tradução minha do seguinte trecho: "El grado de densidade del entramado conceptual permite realizar afirmaciones fundadas acerca del nível de especialidade del texto."

(*función*), a situação em que se insere, no que diz respeito à comunidade linguística, e o procedimento, em que há a seleção dos temas e sua tessitura.

Neste trabalho, a relevância empírica das análises realizadas pela estudiosa argentina se dá pelo fato de haver aqui a preocupação com o aprendiz de inglês em nível técnico, o qual costuma, cotidianamente, ler textos que se encaixam nos perfis de artigos de jornal e de divulgação científica. Entretanto, quando se trata de estudar conteúdos das disciplinas técnico-profissionais, estes alunos devem ler artigos científicos, os quais têm um alto grau de especialidade e, portanto, complexidade. Assim, compreender esses níveis de sofisticação é importante para inferir os graus de dificuldade em se tratando de leitura, especialmente em contexto de língua estrangeira. Em outras palavras, trata-se de leitores brasileiros semiespecialistas, que estão em formação acadêmica, os quais devem ler textos anglófonos voltados para o público de especialistas.

O ponto de convergência entre as duas correntes terminológicas aqui empregadas é a maneira de olhar o texto especializado, levando em consideração a sua função comunicativa e o contexto que cerca os termos. Como neste projeto há a preocupação de se levantar as colocações especializadas da Ciência da Computação para fins de ensino de inglês instrumental, a ótica sob a qual se levantará e analisará os dados leva em consideração tanto a língua especializada em uso, quanto o alto grau de especialidade do gênero textual aqui estudado: o artigo científico.

2.4 Colocações e linguagens especializadas

Conforme já foi mencionado anteriormente, a LC está presente nos estudos que se debruçam sobre o uso das línguas. Esse uso refere-se a estruturas mais ou menos recorrentes, bem como seus contextos. Salienta-se, portanto, que tal abordagem observa tanto a frequência quanto as situações de ocorrência. Algo bastante relevante nos textos são as convencionalidades a que ele se submete, devido aos propósitos linguísticos almejados. Na

comunidade discursiva (SWALES, 1990) acadêmica, as colocações são bastante recorrentes, contendo ou não termos em sua composição. Assim, LC e Terminologia, mais especificamente quanto ao emprego dessas convenções, exige uma parceria dessas duas áreas (ZILIO, 2010).

O significado de uma palavra "é sempre contextual" (FIRTH, 1957, p. 7). Por isso, propõe-se aqui estudar as colocações sob a ótica de que as palavras se constituem de acordo com a companhia que elas têm. As coocorrências podem ser arbitrárias, mas "se um de seus componentes for alterado, poderá haver ruído na comunicação" (SANTOS, 2010, p.100).

No âmbito dos estudos do léxico, as fraseologias são definidas como estruturações linguísticas estereotipadas, que levam "a uma interpretação semântica independente dos sentidos estritos dos constituintes" (KRIEGER; FINATTO, 2004, p. 84). Incluem-se nesse universo as frases feitas, locuções verbais e nominais, expressões idiomáticas, provérbios e colocações. Elas "costumam expressar um significado que não é deduzível das partes dessa combinação" (KRIEGER; FINATTO, 2004, p. 85) de elementos lexicais, comportando traços cognitivos. Em se tratando de língua geral, o que se chama aqui de fraseologias tem uma grande diversidade denominativa, de acordo com o conceito adotado para atender os objetivos de cada pesquisa, assim, o que alguns autores nomeiam de unidades fraseológicas, outros podem denominar colocação, coocorrente ou, ainda, fraseologismo (BEVILACQUA, 2004). Já no âmbito do texto especializado, essas construções são estabelecidas no âmbito das convencionalidades, e não mais das idiomaticidades (TAGNIN, 2013), posto que são determinadas por sua comunidade discursiva.

As colocações, por sua vez, distinguem-se das combinatórias livres, nas quais as palavras se combinam por mera afinidade. Nas colocações, um não nativo pode entender as palavras que as compõem, mas não saberá usá-las e/ou reproduzi-las, não sendo essa tarefa algo automático. Esse aspecto relaciona-se com as próprias características das colocações, entre elas, a arbitrariedade/imprevisibilidade de escolha dos elementos que a compõem e que, pelo seu uso frequente, acabam se fixando. Dentro de uma colocação, o

status de cada uma de suas partes não é igual. Hausmann (1990) denomina base[5] (palavra-chave) o componente que não modifica a identificação do que é caracterizado; *collocatif* (colocado) é o componente que só recebe seu significado semântico através da construção da colocação em si. Segundo Cop (1988), ainda, uma colocação empregada corretamente, por falantes *quasi fluent* ou nativos possui palavras que se atraem como polos negativos: +A → ← B-.

Sinclair (1990) propõe dois tipos de colocações. As *downward collocations* são aquelas em que a palavra central A, que o pesquisador chamava de nó (*node*), tem uma frequência maior que seu colocado B. Isso ocorre, por exemplo, quando se utiliza *make a decision* (variante norte-americana) e *take a decision* (variante britânica). Quando a posição *A*, entretanto, é ocupada por um colocado e *B* é um nó, ocorre uma *upward collocation*. Um exemplo seria nas situações em que se utiliza a terceira pessoa do singular, como em *She makes*. Essa diferenciação era estabelecida pelo renomado linguista de *corpus*, pois ele acreditava que uma *downward collocation* proporciona a análise semântica de uma palavra e ocorre com mais frequência. Dessa forma, as *upward collocations*, por outro lado, são estatisticamente menos recorrentes, e seus componentes lexicais "tendem a ser elementos de construções gramaticais ou superordenados" (SINCLAIR, 1990, p. 116)[6].

Para aprendizes, que se configuram como "falantes ingênuos" (FILLMORE, 1979 apud ALVES, TAGNIN, 2011; TAGNIN, 2013), conhecer como uma língua funciona é primordial. Não basta aprender palavras isoladamente, mas compreender como elas se comportam. "Isso significa dizer que a falta de palavras [isoladas] não introduz apenas lacunas de significados, mas de blocos de informações, uma vez que as palavras trazem consigo instruções de como devem se relacionar entre si" (SCARAMUCCI, 1997). Essa é uma das dificuldades enfrentadas por alunos, a qual se encontra no âmbito das convencionalidades. Essas ocorrem em três níveis linguísticos: sintático, semântico e pragmático (Figura 1).

[5] O texto original é em francês e utiliza as nomenclaturas tal qual citadas em itálico neste artigo.
[6] Tradução minha

O nível sintático trata da combinabilidade dos elementos no que diz respeito à sua ordem e gramaticalidade como, por exemplo, quando se utiliza "velho" coocorrendo com "gagá" formando "velho gagá". O nível semântico engloba a "relação não motivada entre uma expressão e seu significado" (TAGNIN, 2013, p. 26), como quando se diz "chutar o balde" com o sentido de desistir ou "*kick the bucket*", significando morrer. Já o nível pragmático está relacionado com o uso da língua em situações específicas que determinam o que os falantes podem empregar. Isso ocorre, por exemplo, em velórios no qual se pode dizer ao parente do falecido "Meus pêsames" ou simplesmente dar um aperto de mão, um abraço ou um aceno com a cabeça.

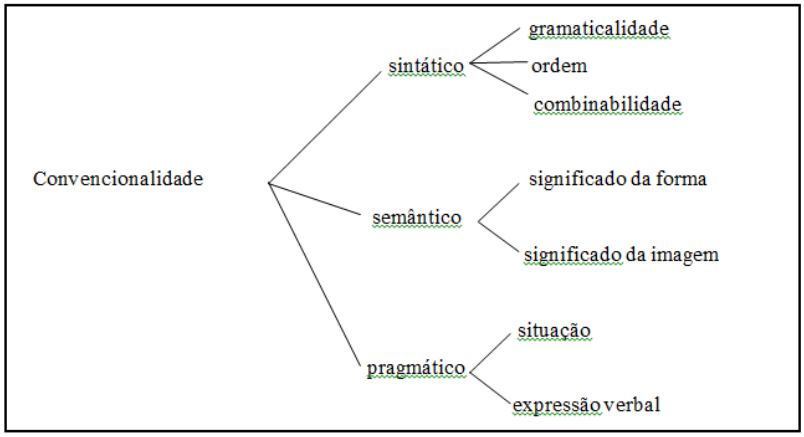

Figura 1 - Níveis de convencionalidade (TAGNIN, 2013, p. 27).

Ao se deparar com todas estas facetas das convencionalidades lexicais, enxergam-se as linguagens especializadas como sublinguagens. Para compreender melhor esse conceito, Hoffmann (2004, p.80) define:

> Uma sublinguagem é um sistema parcial ou um subsistema da linguagem que se atualiza nos textos dos âmbitos comunicativos especializados. Pode-se também dizer: a sublinguagem é um recorte de elementos linguísticos e de suas relações estabelecidas em textos de uma temática limitada.

Logo, a especificidade das linguagens especializadas não se expressa somente através do léxico, mas principalmente por "categorias gramaticais, construções lexicais e outras estruturas textuais" (HOFFMANN, 2004, p. 81), as quais podem ser comprovadas através de métodos estatísticos. Ainda assim,

no que esse autor nomeia de vocabulário especializado, há a predominância de substantivos e adjetivos em detrimento dos verbos. Isso se dá porque no texto especializado, tem-se que designar objetos e manifestações da atividade técnico-científica. Assim, é somente no texto como um todo que se pode explicar melhor o uso especializado de estruturas linguísticas, as quais se inserem em um conjunto de escolhas e convenções que estabelecem um "modo de dizer" (FINATTO; EVERS; ALLE, 2010, p. 155).

Muito embora existam diferentes formas de categorizar as colocações, aqui se optou por adotar a tipologia de Tagnin (2013). Ainda que essa trate da língua geral, e neste estudo o foco esteja nas linguagens especializadas, como tal autora tem grande preocupação com o ensino-aprendizagem de língua inglesa, justifica-se tal aporte teórico. Reitera-se, contudo, que o olhar que se deve ter com relação às ocorrências do *corpus* aqui analisado não dever perder de vista os aspectos estabelecidos por Hoffmann (2004).

2.4.1 Colocações especializadas, aprendizes e inglês instrumental.

Para fins desta proposta, utiliza-se o termo colocação para definir o objeto de estudo, tendo-o como convencionalidade lexical, que pode ser apresentado através da combinatória de substantivos, adjetivos, verbos ou advérbios (TAGNIN; VALE, 2008; TAGNIN, 2013). Em exemplo seria *learning object material(s),* que está convencionalizada em Ciência da Computação, sendo assim empregada em artigos científicos. Como se destina a estudar tal gênero, é relevante trazer a diferenciação que L'Homme e Bertrand (2000, p. 498) fazem das colocações da língua geral e da especializada: "Colocações são convenções dentro de uma determinada comunidade linguística; as combinatórias lexicais especializadas são convenções dentro de um grupo de especialistas".

Essas autoras colocam, ainda, que tanto dizer colocações especializadas quanto combinatórias léxicas especializadas atende a uma mesma comunidade de especialistas. Zilio (2012) observou as combinatórias

recorrentes na área de Cardiologia em corpora de artigos científicos em português e alemão, obtendo êxito com resultados bem interessantes do ponto de vista terminológico. Assim como esse pesquisador, aqui se utilizará a denominação **colocações especializadas** (CEs) em contraste a simplesmente **colocação**.

Em relação à importância das colocações no ensino de inglês instrumental, Scaramucci (1997) verificou que alunos de graduação de uma universidade pública tinham o vocabulário como seu maior problema para ler textos acadêmicos. Além disso, os discentes expressaram que tinham as classes gramaticais como um aspecto associativo “mais fácil” e o os padrões colocacionais como “o mais difícil”.

Para não nativos, seja em situações cotidianas, acadêmicas ou profissionais, a importância de se conhecer e estudar colocações implica entender melhor algo que não pode ser inventado ou parafraseado (SANTOS, 2010). Essas convenções não podem ser vistas como algo com regularidade sintática ou semântica, uma vez que aprendizes devem adquiri-las partindo do pressuposto de que elas são imprevisíveis e arbitrárias (HAUSSMAN, 1990; L’HOMME; BERTRAND, 2000). Em se tratando de textos especializados, as colocações especializadas produzem convenções terminológicas que devem ser apreendidas pelo público de uma determinada área. Nisso reside a relevância de se produzir compilações de CEs de todas as áreas e subáreas do conhecimento. Pesquisadores e estudantes dos mais diversos níveis de proficiência em inglês devem se comunicar de maneira produtiva em seus campos de estudo, divulgando e trocando informações.

3 Materiais e métodos

Tendo como objeto de estudo as CEs presentes em artigos científicos da área de Ciência da Computação, este piloto teve como *corpus* aquele que foi construído por Possamai (2004) em sua dissertação de mestrado. O mesmo contém 112 artigos em inglês de 6 periódicos diferentes, contemplando

divulgação de pesquisas em: Inteligência Artificial, Ensino e Tecnologia, Educação à distância, Robótica, Informática na Educação e Engenharia de Software. Totalizando 836.914 palavras, o *corpus* é considerado de tamanho médio (BERBER SARDINHA, 2004). Com a impossibilidade de verificar a(s) nacionalidade(s) do(s) autor(es) dos textos, a referida pesquisadora foi cautelosa em selecionar artigos provenientes de universidades anglófonas.

Para levantamento e análise do objeto de estudo, utilizou-se a ferramenta AntConc (ANTHONY, 2005)[7]. Primeiramente, obteve-se uma lista das palavras do *corpus* e sua frequência (*wordlist*). Essa trouxe dados relevantes para estabelecer a tomada de decisões quanto ao que seria ou não analisado. Havia várias palavras com uma frequência alta, mas quando verificado seus contextos (função *concordance*) de ocorrência, tal observação não apresentava a constituição de colocações especializadas. Isso ocorreu, porque não foram utilizadas *stopwords*, que são aquelas palavras com as quais se configura a ferramenta para que sejam entendidas como "ruído", ou seja, elas não devem constar da lista geral. Aqui se optou por não desperdiçar tal tipo de dado, aproveitando-se a experiência estabelecida por outros estudos realizados com colocações em textos especializados (SANTOS, 2010; ZILIO, 2012), que não procediam ao descarte imediato, analisando a lista de ocorrências unitárias com bastante atenção. Dessa maneira, somente foi possível observar uma palavra dotada de um sentido mais completo na 40a colocação do ranqueamento: *student*. Ainda assim, esse vocábulo não demonstrou ser uma base para CEs, conforme será demonstrado na próxima seção.

A fim de extrair listas de colocações e poder observar o comportamento estatístico e textual de seus colocados, utilizou-se tanto as funções de levantamento de n-gramas quanto de *clusters* do AntConc, quanto as funções da ferramenta *Collocation Extract* 3.04[8]. Essa última, além de gerar uma *wordlist* que possibilita estudar todas as palavras que ocorreram em forma de

[7] A ferramenta pode ser baixada gratuitamente em <http://www.antlab.sci.waseda.ac.jp/software.html>.
[8] http://pioneer.chula.ac.th/~awirote/colloc/

bigramas, independente de formarem colocações ou não, tendo alguma palavra-chave a ser investigada, a ferramenta também extrai listas de colocações formadas por duas ou mais palavras, com um refinamento maior. Outro recurso de cunho linguístico é a utilização do concordanciador para visualizar os contextos das colocações, e até mesmo remeter ao texto fonte de cada uma delas. Uma desvantagem, entretanto, que essa ferramenta apresenta é que ela não funciona plenamente no sistema Windows 7, por isso é necessário ter uma máquina que tenha o Windows XP. Salienta-se que as duas ferramentas computacionais mencionadas foram empregadas de forma complementar, ou seja, como o critério de frequência é algo relevante, utilizaram-se as listas provenientes dos dois recursos, a fim de subsidiar a análise manual.

Observou-se, ainda, se a CE estava presente em mais de dois artigos, pois se verificou que há padrões lexicais inerentes ao(s) autor(es) ou à especificidade do estudo realizado. Um exemplo é *video-based learning object*, que teve 13 ocorrências no *corpus*, mas todas elas em um único artigo científico.

Como este estudo ainda é incipiente, ratifica-se que o mesmo se constitui como uma proposta, conforme anuncia o título do presente artigo.

4 Resultados

Em virtude dos domínios dos periódicos que compõem o *corpus* de Possamai (2004), percebeu-se haver muitas ocorrências de colocações do tipo Adjetivo + Substantivo com a palavra-chave *learning*. Além disso, obtiveram-se várias colocações que têm em sua composição *learning object*. Definiu-se, assim, o uso de alguns símbolos para indicar convencionalidades recorrentes. Quando a combinatória ocorre somente com o núcleo mais um colocado (bigrama) ou tem algum colocado adicional (trigrama), optou-se por utilizar [*]. Já

quando a colocação se constitui como um n-grama[9], mas o primeiro ou último colocado é variável, utilizou-se apenas o símbolo * para indicar tal fenômeno.

Na Tabela 1, é possível verificar 1.428 CEs que foram levantadas neste estudo. Com certeza, pesquisando o *corpus* por mais tempo, muitas outras ocorrências relevantes para a área de Ciência da Computação aparecerão. Contudo, mesmo de forma incipiente, já é possível levantar dados convincentes acerca do uso de algumas colocações. Percebeu-se que as CEs observadas têm bastante vínculo com a relação empírica entre Computação e Ensino. Assim, palavras como *education* e *learning* se demonstraram bastante recorrentes.

Colocação especializada	**Número de ocorrências**
learning object *	365
* learning	387
[*] autonomous systems	170
object oriented [*]	108
higher education [*]	87
social robots	86
* information technology	74
natural language	37
computer mediated learning	30
control system	26
[*] acess to information *	25
management systems	22
online teaching and learning support	11
	1428

Tabela 1 - Colocações especializadas no *corpus* Possamai

Outro dado verificado com bastante frequência foi o de colocações especializadas compostas por learning object acrescido de um terceiro elemento. Elas se apresentaram em 15 possibilidades diferentes de combinatórias, totalizando 338 ocorrências.

[9] Um n-grama é uma ocorrência lexical que pode ser unitária (unigrama), binária (bigrama) ou com mais elementos (trigrama, tetragrama, etc).

Colocação especializada	Número de ocorrências
learning object interactions	8
learning object principles	7
learning object program	9
learning object design	10
learning object theory	10
learning object resources	13
learning object effectiveness	11
learning object systems	17
learning object environment	75
learning object community	19
learning object communities	49
learning object materials	30
learning object process	29
learning object strategies	27
learning object technology	24
	338

Tabela 2 - Ocorrências de colocações com learning object

Pôde-se observar, ademais, outros contextos para *learning object*, que não estão listados na tabela, por possuírem baixa frequência e por estarem somente presentes em somente uma das fontes do *corpus*, mas que trazem dados interessantes: essa combinatória também tem como colocados *repository, repositories, paradigm e aggregation* (Figura 1). Quando o *corpus* for aumentado, futuramente, será importante observar se tais colocações persistem.

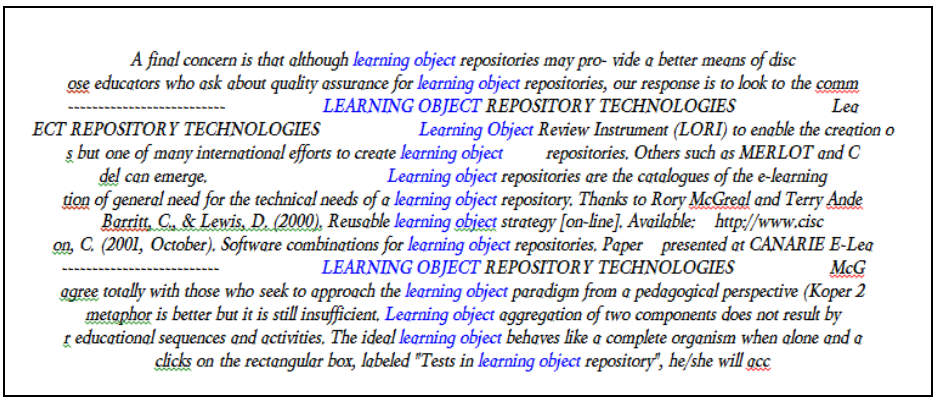

Figura 1 - Ocorrências de colocações com learning object com outros colocados

Muito embora 25 ocorrências não sejam numericamente tão elevado, se comparadas com outras listadas nas tabelas, observou-se que as CEs constituídas por *access to information* (Figura 2) apresentaram diversas composições. Há casos em que são iniciadas por *low* ou *high*. A posposição por *technology* pode ocorrer no singular ou no plural. Isso traz o alerta de, no momento da análise manual, não se ater exclusivamente às colocações compostas por bigramas, conforme a fórmula estabelecida por alguns autores, mas seguir investigando os outros colocados circunjacentes a uma palavra-chave.

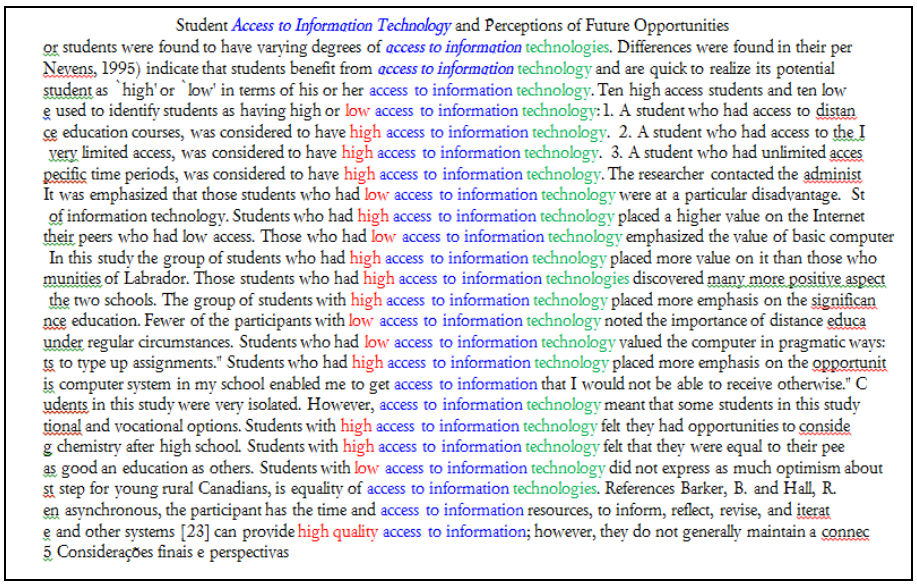
Student *Access to Information Technology* and Perceptions of Future Opportunities
or students were found to have varying degrees of *access to information* technologies. Differences were found in their per
Nevens, 1995) indicate that students benefit from *access to information* technology and are quick to realize its potential
student as `high' or `low' in terms of his or her access to information technology. Ten high access students and ten low
e used to identify students as having high or low access to information technology: 1. A student who had access to distan
ce education courses, was considered to have high access to information technology. 2. A student who had access to the I
very limited access, was considered to have high access to information technology. 3. A student who had unlimited acces
pecific time periods, was considered to have high access to information technology. The researcher contacted the administ
It was emphasized that those students who had low access to information technology were at a particular disadvantage. St
of information technology. Students who had high access to information technology placed a higher value on the Internet
their peers who had low access. Those who had low access to information technology emphasized the value of basic computer
In this study the group of students who had high access to information technology placed more value on it than those who
munities of Labrador. Those students who had high access to information technologies discovered many more positive aspect
the two schools. The group of students with high access to information technology placed more emphasis on the significan
nce education. Fewer of the participants with low access to information technology noted the importance of distance educa
under regular circumstances. Students who had low access to information technology valued the computer in pragmatic ways:
ts to type up assignments." Students who had high access to information technology placed more emphasis on the opportunit
is computer system in my school enabled me to get access to information that I would not be able to receive otherwise." C
udents in this study were very isolated. However, access to information technology meant that some students in this study
tional and vocational options. Students with high access to information technology felt they had opportunities to conside
g chemistry after high school. Students with high access to information technology felt that they were equal to their pee
as good an education as others. Students with low access to information technology did not express as much optimism about
st step for young rural Canadians, is equality of access to information technologies. References Barker, B. and Hall, R.
en asynchronous, the participant has the time and access to information resources, to inform, reflect, revise, and iterat
e and other systems [23] can provide high quality access to information; however, they do not generally maintain a connec
5 Considerações finais e perspectivas

Figura 2 - Ocorrências de colocações compostas por access to information

Conclusões

Entende-se que ainda pode ser melhorada e aprimorada a presente proposta, mas percebe-se também que o caminho que está sendo trilhado parece promissor. Aumentando as temáticas de periódicos do *corpus*, bem como o número de palavras, seguramente mais CEs poderão ser levantadas e analisadas futuramente. Além disso, é necessário atualizar o *corpus*, uma vez que o mesmo contém artigos somente até o ano de 2003. Sabe-se que a área de Ciência da Computação tem rápidos avanços, o que pode acarretar mudanças no léxico empregado.

Ainda há muito a ser explorado no horizonte deste estudo, o que o torna ainda mais fascinante. As investigações iniciais do *corpus* trouxeram dados que poderão auxiliar leitores não nativos de artigos científicos da área de Ciência da Computação a compreender melhor as combinatórias de palavras e aprimorar tanto sua competência lexical quanto sua formação profissional.

Referências Bibliográficas

ALMEIDA, Gladis M. B. A teoria comunicativa da terminologia e a sua prática. **Alfa**, Araraquara/SP 50(2), 2006, p. 85-101.

ALUÍSIO, Sandra e ALMEIDA; Gladis M. de B. O que é e como se constrói um córpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Calidoscópio**, São Leopoldo/RS, v. 4, n. 3, p. 155-177, 2006.

ALUÍSIO, Sandra. **Ferramentas para Auxiliar a Escrita de Artigos Científicos em Inglês como língua estrangeira**. 1995. Tese (Doutorado em Física) - Instituto de Física, Universidade de São Paulo, São Carlos/SP.

ALVES, Fábio; TAGNIN, Stella E. O. Corpora e Ensino de Tradução: o papel do automonitoramento e da conscientização cognitivo-discursiva no processo de aprendizagem de tradutores novatos. In: VIANA, Vander; TAGNIN, Stella. **Corpora no Ensino de Línguas Estrangeiras**. São Paulo: HUB Editorial, 2011, p. 189-203.

ANTHONY, Laurence. AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. *In*: **Proceedings of International Professional Communication Conference IPCC**, 2005, pp. 729-737.

BAKHTIN, Mikhail. **Estética da Criação Verbal**. São Paulo: Martins Fontes, 1997.

BEVILACQUA, Cleci. **Unidades Fraseológicas Especializadas Eventivas: descripción y reglas de formación en el ámbito de la energía solar**. 2004. Tese (Doutorado em Linguística Aplicada - Léxico) - Instituto Universitário de Linguística Aplicada (IULA), Universidade Pompeu Fabra, Barcelona, 2004.

CABRÉ, Maria Teresa. Textos especializados y unidades de conocimiento: metodología y tipologización. *In*: PALACIOS, Joaquín García; FUENTES, Maria Teresa (Eds.). **Texto, terminología y traducción**. Salamanca: Ediciones Almar, 2002, p. 15-36.

Censo da Educação Superior: 2011 - Resumo Técnico. Brasília: Instituto Nacional de Estudos e Pesquisa Anísio Teixeira, 2013a.

Censo Escolar da Educação Básica: 2012 - Resumo Técnico. Brasília: Instituto Nacional de Estudos e Pesquisa Anísio Teixeira, 2013b.

CIAPUSCIO, Guiomar. El término en los textos: una propuesta integradora para el análisis de la variación conceptual. **Actas del RITERM**, Havana, 1998.

COP, Margaret. The function of collocations in dictionaries. *In*: **Euralex Proceedings**. Budapest, Akadémiai Kédó, 1990, p. 35-46.

FINATTO, Maria José. Estudos sobre linguagens e textos científicos e técnicos: o que é uma terminologia textual? **Encontro do VIII CELSUL - Círculo de Estudos Linguísticos do Sul**. Porto Alegre/RS, Universidade Federal do Rio Grande do Sul, 2008.

FINATTO, Maria José; EVERS, Aline; ALLE, Cybele Margareth. Para além das terminologias: estudos de convencionalidade em linguagens científicas. *In*: PERNA, Cristina Lopes; DELGADO, Heloísa Koch; FINATTO, Maria José (Orgs.). **Linguagens Especializadas em Corpora** - modos de dizer e interfaces de pesquisa. Porto Alegre: EDIPUCRS, 2010, p. 152-182.

FIRTH, J. R. **Papers in Linguistics** 1934-1951. London: Oxford University Press, 1957.

GENOVES JR., Luiz Carlos. **Avaliação Automática de qualidade de escrita de Resumos científicos em Inglês**. 2007. 146 f. Dissertação (Mestrado em Computação) - Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos/SP.

HALLIDAY, M.; ANGUS, M.; STREVENS, P. **The Linguistics Sciences and Language Teaching**. London: Longman, Green and C., 1965.

HAUSSMANN, F. J. Le dictionnaire de collocations. *In*: HAUSSMAN, F. J. [et al.]. **An international encyclopedia of lexicography**. V. 1. Berlin, New York: Walter de Gruyter, 1990, p. 1010-1019.

HOFFMANN, Lothar. Conceitos básicos da Linguística das Linguagens Especializadas. Tradução de Maria José B. Finatto. **Cadernos de Tradução**, Porto Alegre/RS, n. 17, outubro-dezembro, 2004, p. 79-90.

HUTCHINSON, Tom; WATERS, Alan. **English for Specific Purposes**. Cambridge University Press, 1987.

KRIEGER, Maria da Graça; FINATTO, Maria José B. **Introdução à Terminologia:** teoria e prática. São Paulo: Contexto, 2004.

KOCH, Ingedore. **O texto e a construção dos sentidos**. São Paulo: Contexto, 2005.

L'HOMME, Marie-Claude; BERTRAND, Claudine. Specialized Lexical Combinations: Should they be described as Collocations or in Terms of Selectional Restrictions? *In*: **Preceedings of Euralex 2000**, Sttutgard-Germany August 8th-12th.

MARCUSCHI, Luiz Antônio. Gêneros Textuais: definição e funcionalidade. *In*: DIONÍSIO, Ângela Paiva; MACHADO, Anna Rachel; BEZERRA, Maria Auxiliadora. (Orgs.). **Gêneros Textuais & Ensino**. Rio de Janeiro: Editora Lucerna, 2002.

McENERY, Tony and WILSON, Andrew. **Corpus Linguistics**. Edinburgh Textbooks in Empirical Linguistics, 1997.

NUNES, Maria da Graça V. O PLN: para quê e para quem? **I Escola Brasileira de Linguística Computacional (EBraLC)**, São Paulo/SP, Universidade de São Paulo, 2008.

NUNES, Maria da Graça *et al*. **Introdução ao Processamento das Línguas Naturais**. Notas Didáticas do Instituto de Matemática Computacional e Computação, São Carlos/SP, USP, 1999.

OLIVEIRA, Lúcia Pacheco. Linguística de *Corpus*: teoria, interfaces e aplicações. **Matraga**, Rio de Janeiro/RJ, v. 16, n.24, jan.-jun 2009, p. 48-76.

POSSAMAI, Viviane. **Marcadores textuais do artigo científico em comparação português e inglês - um estudo sob a perspectiva da tradução**. 2004. 165 f. Dissertação (Mestrado em Teorias do Léxico e do Discurso) - Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre.

SANTOS, Andrea. Análise de colocações adverbiais em inglês para negócios. *In*: **Corpora no Ensino de Línguas Estrangeiras**. VIANA, Vander; TAGNIN, Stella. São Paulo: Hub Editorial, 2010, p. 97-136.

SARDINHA, Tony Berber. **Linguística de Corpus**. São Paulo: Manole, 2004.

SARDINHA, Tony Berber. **Linguística de *Corpus***: Histórico e Problemática. **D.E.L.T.A.**, v. 16, n. 2, 2000, p. 323-367.

SCARAMUCCI, Matilde. A competência lexical de alunos universitários aprendendo a ler em inglês como língua estrangeira. **D.E.L.T.A**., v. 13, n. 2, São Paulo, Agosto 1997.

SWALES, John M. **Genre Analysis** - English in Academic and Research Settings. Cambridge University Press, 1990.

TAGNIN, Stella. **Convencionalidade e Linguística de Corpus**: sua relevância para a escrita científica em língua inglesa. Disponível em: <http://www.escritacientifica.sc.usp.br/wp-content/uploads/FFLCH-USP++Convencionalidade+e+Linguistica+de+Corpus.pdf>. Acesso em: 01 out. 2012.

TAGNIN, Stella. **O jeito que a gente diz** - combinações consagradas em inglês e português. Baureri/SP: Disal, 2013.

TAGNIN, Stella e VALE, Oto. **Avanços da Linguística de Corpus no Brasil**. São Paulo: Humanitas, 2008.

ZILIO, Leonardo. Colocações especializadas em alemão e português na área de Cardiologia. **TradTerm**, São Paulo/SP, v. 20, dezembro/2012, p. 146-177. Disponível em: <http://revistas.usp.br/tradterm/article/view/49049>. Acesso em: 30 Mai. 2013.

ZILIO, Leonardo. Terminologia Textual e Linguística de Corpus: estudo em parceria. *In*: PERNA, Cristina Lopes; DELGADO, Heloísa Koch; FINATTO, Maria José (Orgs.). **Linguagens Especializadas em Corpora** - modos de dizer e interfaces de pesquisa. Porto Alegre: EDIPUCRS, 2010, p. 152-182.