Trab. Ling. Aplic., Campinas, 50(1): 119-132, Jan./Jun. 2011

A PREPARAÇÃO DE MATERIAL TERMINOLÓGICO EM LÍNGUA INGLESA POR MEIO DE FERRAMENTAS LINGUÍSTICO-COMPUTACIONAIS

PREPARATION OF TERMINOLOGICAL MATERIAL IN ENGLISH BY MEANS OF COMPUTATIONAL LINGUISTIC TOOLS

EDUARDO BATISTA DA SILVA*
MAURIZIO BABINI**

RESUMO: O objetivo desse estudo é demonstrar, por meio de análise quantitativa e qualitativa, a eficácia de ferramentas linguístico-computacionais na seleção de terminologia para a produção de material terminológico. Serão apresentadas duas ferramentas linguístico-computacionais (WordSmith Tools e VocabProfile) e, também, sugestões para que o ensino de termos ofereça resultados práticos. A fundamentação teórico-metodológica recorreu a Barros (2004); Berber Sardinha (2000; 2005); Biderman (2001); Cabré (2007); Cobb (2007); Nation, (2003) e Sinclair (2004). O corpus da pesquisa foi constituído exclusivamente de material escrito na língua inglesa em diversas áreas de especialidade. Os procedimentos de preparação de material terminológico são exemplificados a partir de uma das áreas de especialidades utilizadas nos corpora de pesquisa, as Redes Neurais Artificiais. Os resultados obtidos indicam que a utilização do Wordsmith Tools juntamente com o VocabProfile pode fornecer dados importantes para a pesquisa linguistica.
Palavras-chave: linguística de corpus; terminologia; redes neurais artificiais

ABSTRACT: This paper aims to demonstrate by means of quantitative and qualitative analyses the effectiveness of the linguistic computational tools in selecting terminology for the production of terminological material. Two linguistic computational tools will be introduced (WordSmith Tools e VocabProfile) and also suggestions so as the teaching of terms may offer practical results. The theoretical-methodological approach relies on Barros (2004); Berber Sardinha (2000; 2005); Biderman (2001); Cabré (2007); Cobb (2007); Nation (2003) and Sinclair (2004). The research corpus was made solely of written material in English in several specialty languages. The procedures regarding terminological material preparation are exemplified with one of the specialty fields used in the research corpus, the Artificial Neural Networks. The obtained results indicate that the use of Wordsmith Tools in conjunction with VocabProfile might provide useful data for the linguistic research.
Keywords: corpus linguistics; terminology; artificial neural networks.

1. INTRODUÇÃO

Nos últimos anos, os simpósios da RITerm (Rede Ibero-Americana de Terminologia) têm dedicado atenção ao modo pelo qual a terminologia é abordada nos cursos que se propõem a ensiná-la. Em língua portuguesa, os estudos que tratam dessa problemática ainda são incipientes. Devido à importância de determinadas áreas do conhecimento, especialmente daquelas ligadas a informática, o indivíduo que não domina a terminologia de sua atividade está fadado ao insucesso, principalmente no momento em que travar

* UEG, Quirinópolis (GO), Brasil/Doutorando em Estudos Linguísticos da UNESP, São José do Rio Preto (SP), Brasil . eduardo.silva@ueg.br
** UNESP, São José do Rio Preto (SP), Brasil. maurizio@ibilce.unesp.br

contato com textos de língua inglesa de especialidade. No caso da Inteligência Artificial, torna-se necessário o domínio da terminologia associada. Inúmeros cursos vêm sendo elaborados e ministrados para ensinar os termos, a fraseologia especializada etc. Surgem os cursos de inglês para fins específicos, também conhecidos como English for Specific Purposes (ESP), que, como o próprio nome sugere, estão voltados às necessidades do aprendiz.

Frente à escassez de discussões relacionadas ao ensino de terminologia assistida pelo *WordSmith Tools* (WST) e pelo *VocabProfile* (VP), o presente artigo propõe-se a levantar algumas questões pivotais para as quais a Linguística de Corpus pode fornecer insumos. Curiosamente, grande parte dos pesquisadores linguísticos e dos professores deixa de lançar mão de recursos tecnológicos na prática docente. Existe uma tendência incipiente em associar as pesquisas linguísticas aos aplicativos computacionais. Talvez isso ocorra devido à recente popularização das ferramentas computacionais.

Além do contato direto com a terminologia, partimos de nossa experiência na ministração de cursos preparatórios para provas de admissão em níveis de pós-graduação de uma universidade estadual paulista. Essas provas exigem que o candidato compreenda textos escritos de sua especialidade em língua inglesa e faça exercícios optativos de gramática relacionada ao texto. Trata-se de uma avaliação totalmente escrita, que demanda reconhecimento de vocabulário da língua geral, vocabulário de línguas de especialidades e de estruturas gramaticais.

Vale ressaltar que não é intenção desse trabalho analisar as diferentes concepções e discussões teóricas sobre ESP nem as importantes contribuições elaboradas nas últimas décadas sobre o assunto. O objetivo principal deste trabalho é apresentar possibilidades de pesquisa, no que se refere ao levantamento de terminologia de áreas de especialidade para posterior utilização no ensino de línguas estrangeiras, mais especificamente no ensino de língua inglesa em contextos específicos. Em suma, as preocupações presentes ao longo deste trabalho foram, em primeiro lugar, apresentar ferramentas linguístico-computacionais para professores que precisam ensinar terminologia e, em segundo lugar, oferecer uma possibilidade de pesquisa que proporcione resultados práticos para aqueles que estão envolvidos com a terminologia.

2. PANORAMA ATUAL

A contar pela quantidade de cursos para inglês instrumental disponíveis por todo o país, entende-se que deve haver um público ávido por esse tipo de curso. Desde escolas de idiomas até cursos em universidades, a diversificação impressiona.

Muitos deles comprometem-se a fazer o aprendiz entender a mensagem do texto em pouco tempo. A maioria apela para estratégias de inferência de sentido, a partir de informações pictóricas e/ou infográficas. Isso parece ser a solução para os problemas que dificultam a habilidade da leitura de textos em língua inglesa. Leva-se muito em conta a intuição. Muitos acreditam que, com o curso ESP, o leitor estará capacitado a entender um texto de matemática ou de microbiologia, lançando mão de algumas estratégias de leitura, recorrendo também ao conhecimento prévio sobre o assunto.

A maioria dos professores de inglês para fins específicos parece carecer de informações relacionadas à Linguística de Corpus. Dessa forma, desconhecem ferramentas de tratamento linguístico que provam ser úteis nesse tipo de trabalho. As constatações dessa linha de pesquisa acabam ficando encerradas nos círculos acadêmicos. De forma a disseminar os estudos baseados em corpus, esperamos que esse artigo sirva de subsídio para outras discussões.

Não existe nenhum manual ou livro didático disponível no mercado nacional, pelo menos até o momento, que contemple o ensino de terminologia baseada em corpus. Consequentemente, o professor responsável pelo curso de inglês instrumental ou de inglês para fins específicos, encontra-se em uma situação delicada. Muitas vezes, o professor não é um especialista da área em questão. De fato, espera-se que um professor de língua inglesa domine a língua geral. Após algum tempo, uma pessoa que a domine bem, consegue entender um texto especializado. Para isso, o acompanhamento de um especialista ou uma pesquisa terminológica gera resultados satisfatórios no entendimento desse texto.

No desenvolvimento de material terminológico, saltam aos olhos três questões: 1) Os exemplos produzidos pelo professor serão simples demais?; 2) A terminologia ensinada será puramente intuitiva ou será retirada de algum texto especializado?; e 3) Quais são as estruturas gramaticais mais importantes?

A resposta ao primeiro questionamento varia em função da formação do professor. Os exemplos de frases produzidos por um não-especialista, não-nativo, tendem a ser insatisfatórios devido à falta de conhecimento específico da área e à falta de familiaridade com as estruturas recorrentes daquele domínio. Para quem não é especialista, a utilização de terminologia na própria língua nativa prova ser uma atividade difícil. A dificuldade na produção de exemplos em língua inglesa é aumentada exponencialmente.

Para a segunda pergunta, o fato de a terminologia ser retirada de um texto, ou mesmo de alguns textos de especialidade, não determina de forma inequívoca que os termos mais relevantes estejam presentes. E, por último, algumas estruturas gramaticais, são bastante comuns em textos científicos. Vidal e Cabré (2005), em um estudo com corpus de especialidade em espanhol, identificam algumas estruturas comuns, como por exemplo, o predomínio de pronomes, uso considerável da terceira pessoa do plural, predomínio do presente do indicativo e da voz passiva. Biber (1998) também identifica o uso frequente da voz passiva nos textos de especialidade.

As características citadas acima constituem itens relevantes no desenho de um curso ESP. Nos últimos anos, cada vez mais, o ensino do vocabulário das áreas de especialidade vem ganhando destaque. As pesquisas que envolvem tanto o léxico geral quanto o léxico das áreas de especialidade vêm sendo desenvolvidas por Cabré (1993), Nation (2001; 2003), Barros (2004), dentre outros. Coxhead e Hirsh (2007), por exemplo, elaboraram seis listas contendo termos ligados à ciência. O objetivo das listas é auxiliar no aprendizado da terminologia científica em língua inglesa.

No tocante ao conhecimento lexical em língua inglesa, Scaramucci (1997) avalia estudantes da área de Engenharia Elétrica e Matemática Aplicada que estão aprendendo a ler em inglês como língua estrangeira. Faziam curso de inglês instrumental. A eles são apresentados dois textos pertencentes a domínios distintos. O primeiro versa sobre nascimento de uma galáxia e o outro sobre ecologia. A referida autora diz que 82% dos

sujeitos em sua pesquisa apontam o vocabulário como sendo o maior problema. A pesquisa em questão revela que os sujeitos com mais conhecimento de vocabulário obtiveram melhores resultados de inferência em contexto. Aqui, provavelmente, o vocabulário é entendido como da língua geral, abarcando também aquele das línguas de especialidade.

Um aluno que deseja fazer um curso de inglês instrumental pode ter dois motivos: ser capaz de entender "qualquer" texto em língua inglesa ou entender os textos de sua área de especialidade. Seja no primeiro ou no segundo caso, um planejamento cuidadoso deve ser feito, tendo em mente o que, de fato, é relevante para os aprendizes.

Entende-se que para entrar em contato com a leitura de textos em inglês, as unidades lexicais e as terminológicas tem posição de destaque. Não deixando de atentar para outras estruturas integrantes das línguas de especialidade, Vidal e Cabré (2005) destacam a importância dessa questão, atestando que são necessários estudos descritivos prévios sobre a combinatória léxica especializada, permitindo conhecer quais aspectos são os mais problemáticos e, desse modo, encaminhar a docência à sua resolução

Com relação ao ensino de vocabulário, Scaramucci diz que "além de evidências esparsas e muitas vezes conflitantes, não se encontram, na literatura sobre vocabulário, indicações mais precisas de como seria um ensino de vocabulário que conduza à compreensão" (SCARAMUCCI, 1997). Mais recentemente, Nation (2001; 2003) fornece algumas possibilidades de como estruturar o aprendizado de vocabulário em língua inglesa, nas habilidades de produção e de recepção de língua estrangeira.

3. A SELEÇÃO DO VOCABULÁRIO

Hoje, com a linguística de corpus, a identificação das palavras ou itens (types) mais importantes em uma área de especialidade apresenta-se como uma importante ferramenta do professor. Desde 1921, quando Thorndyke criou uma das primeiras listas de frequências do inglês norte-americano com o propósito de facilitar o ensino de inglês, outros estudiosos têm se dedicado à mesma empreitada.

Pesquisas de cunho estatístico-lexical indicam que o domínio das 2000 palavras mais comuns da língua geral permite a compreensão de mais de 70% de um texto. Estudos anteriores, e mais recentemente, Cobb (2007) e Nation (2003) trabalham com o mesmo recorte de palavras. Nation salienta que as palavras de alta frequência merecem atenção do professor, do aprendiz e do livro-texto. Sob a mesma orientação em termos de frequência, Biderman (2001), em uma pesquisa com a língua portuguesa, descobriu que 2000 palavras perfaziam 84% de seu corpus.

Os dicionários das editoras, tais quais Cambridge, Longman, Oxford e Collins-Cobuid, restringem a quantidade de palavras que utilizam na definição dos verbetes. O Cambridge Dictionary of American English lança mão de 2000 palavras para definir pouco mais de 84.000 verbetes.

O monitoramento do léxico, que já é tido como indispensável na prática lexicográfica para aprendizes da língua inglesa, ainda não ganhou o devido espaço na seara dos estudos terminológicos. Via de regra, os materiais para aprendizes são desenvolvidos com base na intuição dos autores. Textos que os alunos trazem ou textos que os próprios professores

acreditam ser importantes são trabalhados em aula. Com isso, gasta-se tempo com estruturas gramaticais raras e palavras que possuem baixíssima frequência naquela área do conhecimento. Seguindo a intuição, palavras e mais palavras são consideradas essenciais. A todo momento, os termos elencados e retirados daqueles textos sobre o assunto em questão são destacados como sendo importantes. Ora, alguns poucos textos não conseguem fornecer insumo terminológico de maneira adequada. Parece não existir uma preocupação formal com a utilização das palavras mais comuns a partir de observações estatístico-computacionais.

4. FERRAMENTAS PARA A PESQUISA LINGUÍSTICO-ESTATÍSTICA

A seguir, são apresentadas duas ferramentas que podem ser utilizadas na preparação de material terminológico. Trata-se do software *WordSmith Tools* e do software *VocabProfile*.

4.1. Wordsmith tools

O advento da ferramenta *Wordsmith Tools* contribuiu para impulsionar a popularização de estudos baseados em corpus. Com os recursos que o WST e os computadores pessoais disponibilizam ao usuário, está ao alcance de professores, terminólogos, linguistas, tradutores etc, a manipulação de corpora das linguagens conhecidas como técnicas. O desenvolvimento, bem como o planejamento, de material voltado para a preparação de conteúdo terminológico, satisfaz uma gama de indivíduos em busca de tal recurso. A partir das informações reveladas pelo WST, outras reflexões são possíveis e adequações são sugeridas para que o material possa ser trabalhado de forma a proporcionar melhores resultados.

Com o WST, a identificação das palavras mais frequentes em um corpus é executada em segundos. Para a captura dos termos mais frequentes, o recurso das palavras-chave auxilia. Uma vez que os termos aparecem diversas vezes no texto científico, o programa WST faz a comparação com o corpus de referência e fornece uma lista com os itens (types), que possivelmente se caracterizam como termos. Para que o pesquisador ateste o uso de determinado termo, o concordanceador revela quais palavras são usadas com determinado item. Pode-se configurar o programa para a visualização da quantidade de palavras à esquerda e à adireita do termo selecionado. Pode-se, dessa forma, testemunhar o comportamento do candidato a termo em seu meio natural.

4. 2. O Vocabprofile

O *Vocabprofile*, v. 2.9, é um programa para análise linguística que divide o texto em várias faixas de frequência. Foi desenvolvido por Tom Cobb, professor da Universidade de Québec. Pode-se utilizar o *VocabProfile* gratuitamente diretamente de sua homepage ou efetuando o download no computador.

Ao inserir um texto na janela de consulta, o VP executa a análise e retorna uma página de resultados. Nessa página, o texto é recortado em partes que são indicadas da seguinte maneira: *K1 words*: As primeiras 1000 palavras mais frequentes da língua inglesa. O VP chama essas palavras de 1K e deixa todas as palavras dessa faixa com a cor azul. Variam bastante. Podem constituir por volta de 70% de um texto científico. *K2 words:* As próximas 1000 palavras mais frequentes da língua inglesa. Elas aparecem no VP na cor verde. O domínio dessa faixa aumenta aproximadamente em 5% o entendimento das palavras de um texto. *AWL:* As palavras acadêmicas. São as palavras utilizadas em textos científicos de diversas áreas de especialidades. O conhecimento dessas palavras acadêmicas pode aumentar o entendimento de um texto por volta de 10%. *OFF-list:* As palavras que não se encontram em nenhuma das listas acima. Aqui, pode-se encontrar todos os nomes próprios, possíveis termos, ou ainda as palavras que se encontram acima das 2000 mais comuns da língua inglesa.

Figura 1 - Screenshot do VocabProfile online.

WEB VP OUTPUT FOR FILE: 1.txt

Words recategorized user as 1k items (proper nouns etc): NOME (total 0 tokens)

	Familes	Types	Tokens	Percent
K1 Words (1-1000):	299	421	**3160**	**59,65%**
Function:	--	--	(1537)	(29,95%)
Content			(1573)	(29,69%)
> Anglo-Sax	--	--	(200)	(5,64%)
= Not Greco-Lat/Fr Cog:				
K2 Words (1001-2000):	53	68	192	3,62%
> Anglo-Sax	--	--	(47%)	(0,89%)
1k+2k			...	(53,27%)
AWL Words (academic):	131	176	**604**	**11,40%**
> Anglo-Sax	--	--	(47)	(0,89%)
Off-List Words	?	331	**758**	**14,31%**
	483?	994	5298	100%

Words in text (tokens):	5298
Different words (types):	994
Type-token ratio:	0,19
Tokens per type:	5,33
Lex density (content words/total)	0,70
Perfaining to onlist only	
Tokens:	3956
Types:	665
Families:	483
Tokens per family:	8,19
Types per family:	1,38
Anglo-Sax Index: (A-Sax tokens + functors / onlist tokens)	50,05%

Greco-Lat/Fr-Cognate Index (Inverse of above) 49,95%

A figura 1 mostra os dados estatísticos de um texto pertencente ao corpus de especialidade das RNA. Apesar de caracterizar-se por possuir alta densidade terminológica, as amostras utilizadas provam que é possível trabalhar com material terminológico de forma prática e isenta de subjetividade.

Com o VP, o pesquisador ou o professor pode selecionar o vocabulário que quer trabalhar. Caso os alunos não dominem o vocabulário básico da área, é possível recorrer às palavras presentes no próprio corpus e desenvolver exercícios variados com a primeira faixa de palavras mais comuns.

Se os alunos já demonstram saber a primeira faixa de palavras, pode-se treinar a segunda faixa. O próximo passo é a prática das palavras acadêmicas. Nation (2003) aponta que as palavras acadêmicas devem ser idealmente estudadas após o domínio das primeiras 2000 palavras.

4.3. COMPARAÇÃO ENTRE O WST E O VP

Comparamos algumas características mais conhecidas do WST e do VP. Conforme indicado na tabela 1, de modo geral, o WST supera o VP em relação à gama de recursos disponíveis. Após a comparação dos 13 itens a seguir, percebe-se que o WST apresenta-se como uma ferramenta mais completa, graças aos recursos mais comumente empregados.

No entanto, a análise qualitativa viabilizada pelo VP é o recurso de maior destaque quando faz-se necessário dividir o texto/corpus em faixas de frequência. Nesse ponto, o VP fornece em poucos segundos o resultado de sua análise computacional.

Tabela 1 - Recursos disponíveis no WST e no VP.

	WST (version 3)	**VP (version 2.9)**
contagem e listagem de itens	👍	👍
contagem e listagem de ocorrências	👍	👍
análise qualitativa		👍
concordância	👍	
clusters	👍	
palavras-chave	👍	
razão type/token	👍	👍
visualização integral do texto		👍
procura avançada	👍	
ajustes	👍	
dados estatísticos	👍	👍
software livre		👍
layout		👍

Tanto o WST quanto o VP auxiliam de maneira satisfatória a pesquisa linguística. A preparação de material terminológico pode ter seu conteúdo desenvolvido com o auxílio dessas duas ferramentas de tratamento linguístico-estatístico.

5. UMA ABORDAGEM BASEADA EM CORPUS

O ponto de partida com relação ao ensino de terminologia baseado em corpus, a nosso ver, seria uma abordagem que privilegia a situação contextual na qual as palavras e/ou termos mais comuns ocorrem. Não se está propondo aqui que o aprendiz simplesmente receba uma lista de palavras mais comuns, acompanhadas dos termos mais frequentes. Uma atitude dessas possivelmente acrescentaria muito pouco ao processo de compreensão do texto. A adequação linguística dos enunciados de acordo com o nível dos aprendizes deve render uma alta porcentagem de compreensão.

Sugerimos a criação de exercícios que explorem o vocabulário com base em produções reais de especialistas. Os exercícios com estruturas recorrentes em textos científicos também contribuem para a familiarização com as mesmas. Os termos podem ser inseridos

já nessa fase. Os aprendizes podem ser expostos a produções contextualizadas, tais quais aquelas que aparecerão em textos seguintes.

> [...] o termo pode ser analisado em seus diferentes aspectos: do ponto de vista do significante e do significado, das relações de sentido que mantém com outros termos (sinônimos, homônimos etc.), de seu valor sociolinguístico (usos, preferências, conotações, processo de banalização etc.) e outros. Os conhecimentos resultantes desses estudos básicos dão sustentação teórica ao trabalho de diversas ciências aplicadas. (BARROS, 2004, p. 40)

Para ilustrar essa proposta, podemos citar a ocorrência de duas estruturas "to be likely to" e "will". Os seis corpora abaixo, a saber Engenharia de Alimentos (EA), Genética (Gen), Matemática (Mat), Microbiologia (Mic), Redes Neurais Artificiais (RNA) e Zoologia (Zoo) atestam a ocorrência dessa estrutura.

Tabela 2 - Ocorrências de estruturas gramaticais nos 6 corpora.

Corpora (ocorrências)	**Ocorrências no corpus (porcentagem no corpus)**	
	to be likely	**will**
EA (1.467.078)	360 (0,025)	1578 (0,025)
Gen (2.163.003)	591 (0,028)	2602 (0,027)
Mat (3.158.610)	124 (0,004)	4062 (0,004)
Mic (2.031.679)	582 (0,029)	2481 (0,029)
RNA (1.457.442)	130 (0,009)	2215 (0,009)
Zoo (3.334.408)	711 (0,021)	3571 (0,021)

A estrutura *to be likely* é bem mais comum nos textos ligados à área da Microbiologia e Genética. Por outro lado, na área da Matemática e das RNA a frequência é bem menor. Essa simples demonstração serve para nortear os desenvolvedores de materiais terminológicos e fazer como que dediquem atenção não apenas aos termos, mas também àquelas estruturas gramaticais que podem ser comuns a todos os demais domínios.

Se existe a pressão do tempo e a dificuldade do aprendiz em memorizar e aprender o máximo de palavras possível, vale a pena sacrificar algumas estruturas menos recorrentes.

A estrutura que se destaca em um corpus de especialidade sendo estudado deve ser enfatizada na prática com os aprendizes. A terminologia associada também pode ser embutida em situações contextualizadas. Vidal e Cabré (2005) acreditam que a terminologia não deve ser ensinada nem aprendida como um conjunto fechado de termos, como um pacote de unidades que têm que ser memorizadas isoladamente.

No recurso denominado concordanceador do WST, a partícula *will* é mostrada exatamente no contexto no qual foi produzida. Como seria pouco prático ler ou tentar retirar manualmente os padrões de uso de todas as 2215 vezes que *will* ocorreu, pode-se utilizar o recurso "clusters" que aparece nessa mesma janela.

Os aglomerados, ou sequência de palavras em seus contextos da área de especialidade poupam o trabalho exaustivo de leitura de todas as ocorrências. Com essa funcionalidade "clusters" do WST, pode-se iniciar a contagem de quais verbos, substantivos ou preposições estão comumente associados ao *will*. Com 65 ocorrências, *we will use* é a aglomeração que se destaca nessa listagem. Em segundo lugar aparece o aglomerado *will be discussed*.

Uma vez que ambos têm origem latina, fica mais fácil apresentá-los aos aprendizes. A terminologia a ser embutida, nesse caso, pode variar bastante. Os aprendizes, nesse caso, até podem participar do processo de montagem de frases. Ou, se o curso não permite essa interação, o professor pode adaptar frases a partir da frequência de outros itens terminológicos mais comuns. Assim, consegue-se viabilizar um material customizado baseado em produções reais.

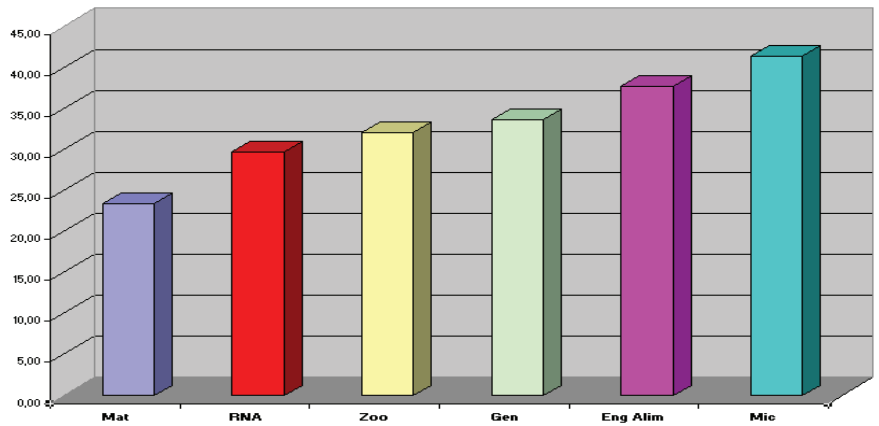
6. A RIQUEZA LEXICAL

Na preparação de material terminológico para aprendizes, mostra-se importante a seleção dos termos mais frequentes. Outrossim, das palavras (tokens) que não são termos, aquelas que ocorrem com alta frequência. Conforme demonstraremos a seguir, diferentes domínios terminológicos apresentam diferente densidade terminológica.

A fim de descobrir se um corpus especializado possui mais termos que outro de tamanho parecido, é possível fazer um cálculo relativamente simples e que oferece valores na maioria das vezes satisfatórios: a razão type-token. A razão type-token é obtida pela divisão do número de itens pelo de ocorrências. No entanto, Gómez (2002) tece algumas considerações sobre esse cálculo e sugere que seja substituído pelo valor K. A fórmula para obtenção desse valor é a seguinte: type=K√token. A margem de erro prevista é de ± 5%.

Seguimos a mesma metodologia de Gómez para realizar um levantamento da densidade lexical de corpora distintos, com tamanhos variados. Pela contagem do K-value, recorremos aos 6 corpora de especialidade, já mencionados anteriormente. Apesar do fato de cada corpus totalizar um número diferente de ocorrências e itens, é possível vislumbrar quais deles possuem uma densidade ou riqueza lexical mais pronunciada, independentemente do tamanho. A Figura 2 exemplifica essas especificidades ao revelar os valores K.

Figura 2 - Densidade lexical de 6 linguagens de especialidade



Depreende-se que no domínio da Matemática, a densidade terminológica é maior. Isso explica-se pelo fato de a matemática ser uma ciência lógica, da qual há maior uso de frases diretas e pouca necessidade de recursos metafóricos ou jogos de linguagem para transmitir a mensagem.

Nosso interesse está voltado especialmente para o corpus das RNA. Procederemos a sua análise, uma vez que não é do escopo deste artigo analisar os demais corpora de especialidade que foram constituídos para a realização deste trabalho.

Nos próximos itens, apresentaremos as etapas necessárias para o levantamento da terminologia no domínio das RNA para seu eventual uso em uma aula de ESP, a saber: uma breve apresentação dessa área de especialidade, constituição do corpus e possíveis aplicações da pesquisa realizada.

7. AS REDES NEURAIS ARTIFICIAIS

As RNA são estruturas capazes de tomar decisões. Após minuciosa análise de padrões e dados previamente inseridos no sistema, as RNA, baseadas em cálculos probabilísticos, apresentam respostas à tarefas das mais variadas. "O que torna essas redes extremamente interessantes aos pesquisadores de muitas áreas é justamente sua capacidade de aprender". (BABINI e MARRANGHELLO, 2007, p. 26)

No que diz respeito às RNA, o termo "camada" encontra algumas definições no Novo dicionário Aurélio século XXI (2004). Dessas, existe uma definição para cada especialidade a seguir: a botânica, a geologia, a eletrônica, a arqueologia e a biologia. Para a física e a astronomia são duas. A geofísica tem três definições para *camada*. O sentido que esse termo possui no domínio das RNA não encontra definição, já que o termo *layer* não foi incluído nesse verbete. Em uma hipotética consulta, o consulente que intenta descobrir o significado de *layer*, nesse caso, não teria suas dúvidas sanadas.

Esse é um exemplo de um item lexical de maior uso na língua geral que torna-se usual em uma língua de especialidade. O contrário também pode acontecer. Vale salientar que uma unidade terminológica pode ocorrer prioritariamente em um único domínio, como é o caso de *backpropagation*.

No entanto, a presença dos termos das RNA nos dicionários ainda é deficiente. Sua terminologia não é abonada nos grandes dicionários em língua inglesa. Um exemplo dessa deficiência é o caso, dentre outros, do termo *backpropagation.*

8. A CONSTITUIÇÃO DE UM CORPUS DE ESPECIALIDADE

O recurso chamado *Keywords* do WST aponta os cinco termos elencados em sua lista como sendo palavras-chave no domínio das RNA. Frente à ausência dos termos das RNA nos dicionários de língua geral, os indivíduos que buscam aprender a terminologia dessa área, precisarão receber um material desenvolvido especialmente para suas necessidades.

Quais são, de fato, as palavras-chave quando torna-se necessária a familiarização linguística com o domínio das RNA, em inglês? Como o aprendiz pode ser exposto a

produções reais dessa área de especialidade, uma vez que nem mesmo os dicionários trazem essas palavras-chave?

Esses questionamentos levam a crer que a constituição de um corpus de especialidade revela-se como fundamental para suprir essa carência.

Somente após análise do corpus de especialidade, é que o professor poderá identificar factualmente o que levará o aprendiz a entender um texto especializado. Caso haja a necessidade de definir os cinco termos acima, o corpus de especialidade pode fornecer informações preciosas relativas ao conceito do termo. Se o objetivo da aula for a tradução dos termos, um trabalho com corpus paralelo em língua portuguesa seria recomendável.

No trabalho com corpus, Sinclair (2004) compara um corpus da ciência da computação com o corpus de língua geral LOB. Ambos com um milhão de ocorrências. Os resultados indicam que a quantidade de itens é 40% menor no texto de especialidade. Isso demonstra que o corpus de especialidade possui menos variedade, mas concentra, por sua vez, mais vocabulário técnico.

A fim de testar o postulado de Sinclair e aplicar nossa proposta para o ensino de terminologia, decidimos constituir um corpus das RNA. O corpus das RNA contabiliza 1.457.442 ocorrências e 35.842 itens. De acordo com a classificação sugerida por Berber-Sardinha (2000), no tocante às dimensões de corpora, trata-se de um corpus médio-grande.

A intenção foi criar um corpus que fosse representativo do domínio das RNA, na língua inglesa. Foram incluídos livros fundamentais e avançados sobre o assunto, guias, websites, tutoriais, apresentações, seminários etc.

Conforme demonstrado na Figura 2, o domínio das RNA, devido a características intrínsicas apresenta menor diversidade lexical, o que sugere maior quantidade de termos.

Por estarem sempre presentes nos textos de especialidade, os termos podem ser caracterizados como palavras-chave. Após estudar uma coletânea de 40 corpora, Berber Sardinha (2005) estima que a quantidade média de palavras-chave seria da ordem de 1472. Alguns corpora alcançam 3000 palavras-chave. Em um corpus de especialidade, as palavras-chave constituem peças fundamentais, já que são representantes do domínio ao qual pertencem. A frequência com que elas ocorrem nos textos especializados (e pouco ou raramente no corpus de referência) justificam a escolha dessas palavras pela importância das mesmas. O recurso keywords do WST geralmente elenca 500 termos, mas pode ser configurado para apresentar um número maior, dependendo da preferência do usuário.

No XIX Seminário Nacional de Inglês Instrumental, Scott (2005) discorre sobre os benefícios que a linguística de corpus pode trazer ao inglês instrumental. Scott lembra que as palavras-chave têm papel importante nesse tipo de atividade.

Além das palavras-chave, Orenha salienta que,

> um dos pontos de convergência entre o ESP e a L.C. é que, da mesma maneira que o ensino de ESP prevê a compreensão dos textos através de exemplos reais, a L.C. é baseada no fato de que a linguagem deve ser estudada através de exemplos reais de uso. (ORENHA, 2004, p. 1038)

A partir dessa seleção, um planejamento mais eficaz de terminologia pode ser levado a cabo. Com relação aos termos das RNA, Silva (2009) indica que apenas 10% dos termos mais frequentes são unidades terminológicas compostas por um termo apenas. Por outro lado, 90% dos termos do levantamento estão estruturadas entre 2 e 5 unidades terminológicas.

9. POSSÍVEIS APLICAÇÕES, EM SALA DE AULA, DA TERMINOLOGIA LEVANTADA

De posse dos termos fornecidos pelo WST ou das palavras destacadas pelo VP, é possível elaborar material destinado ao ensino de terminologia. A fim de contemplar a terminologia e elaborar material baseado em corpus, apresentamos algumas sugestões de possíveis aplicações práticas que podem ser utilizadas em sala de aula para testar habilidades de produção ou de recepção de língua inglesa, desde os níveis mais básicos aos níveis mais avançados.

a) *Palavras-cruzadas*: existem softwares que criam palavras-cruzadas tanto na versão freeware (Criss-Cross Puzzle, da Discovery Education) quanto na versão shareware (Crossword Construction Kit). O exercício pode exigir a tradução dos termos ou, alternativamente, apresentar o conceito dos termos para que os próprios termos sejam incluídos nas linhas ou colunas correspondentes.

b) *Imagens*: os sites que dispõem de imagens dos mais variados tipos podem ser facilmente encontrados em uma consulta a qualquer motor de busca, como o Google, por exemplo. Alguns permitem o download direto da imagem desejada, ao passo que outros solicitam uma inscrição para que o usuário possa efetuar o download. Com diversos planos de pagamento para uso de suas ilustrações, o *iStockphoto* oferece um serviço de qualidade. Uma vez que a grande parte dos termos se enquadra na categoria dos substantivos, exercícios de associação podem ser úteis na memorização e familiarização da terminologia em língua inglesa.

c) *Textos de determinada área de especialidade com lacunas:* as lacunas devem ser preenchidas com termos. Tais termos podem ser previamente fornecidos ou as opções podem ser discutidas no contexto em que ocorrem. O software gratuito *HotPotatoes* possui seis recursos destinados à elaboração de exercícios linguísticos, inclusive um gerador de lacunas nos textos, que pode ser configurado pelo professor.

d) *Frases/parágrafos com variados níveis de dificuldade*: o VP destaca todas as palavras presentes em um texto. Frases/parágrafos podem ser utilizados para checar a compreensão ou interpretação dos termos do trecho selecionado. Para isso, as palavras-chave do WST, que potencialmente se caracterizam como termos, são de grande valia.

e) *Aglomerados* (clusters): o WST destaca *network* como sendo uma palavra-chave no domínio das RNA. O próprio WST também mostra que existem outros itens associados à palavra *network*, por exemplo, *time delay neural network.* Tal recurso propicia maior entendimento das palavras que acompanham outras.

f) *Mapas conceituais*: existem softwares que podem ser usados para estruturar determinada área do conhecimento por meio de sua terminologia. O software *Cmap Tools*, de utilização gratuita, elabora mapas conceituais, facilitando a representação do conhecimento terminológico. Esse recurso pode servir para fixação do conteúdo terminológico obtido pelos aglomerados (clusters) do WST.

10. CONSIDERAÇÕES FINAIS

Ainda que a mente seja uma máquina poderosa na produção de repertórios, deixa a desejar tanto na quantificação quanto na seleção de dados volumosos. Em poucos segundos, um computador consegue tratar milhares de textos e milhões de palavras e apresentar dados confiáveis. Tanto a intuição quanto a subjetividade falham em fornecer uma seleção adequada às necessidades de quem procura o aprendizado de terminologia, seja em língua materna ou estrangeira.

Para o indivíduo que precisa de familiarização com áreas de especialidade, RNA por exemplo, a consulta ao dicionário deixa a desejar. Compreensivelmente, os dicionários de língua geral não conseguem abarcar a terminologia de todas as áreas do conhecimento.

Conforme demonstrado, o uso do WST aliado ao VP possibilita diversas formas de exploração do corpus. Posteriormente, caso o professor ache necessário, outros softwares podem ser utizados para enriquecer o ensino e prática da terminologia. Nossa proposta de manipulação do texto apresenta resultados que, se executados pelo homem, não seriam exatamente fiáveis. O uso da tecnologia e das ferramentas computacionais auxiliam a pesquisa linguística, permitindo o tratamento do texto sob um outro olhar.

Esse artigo pretendeu apresentar que é preciso haver uma seleção terminológica para que os resultados de aprendizado sejam maximizados. O controle de vocabulário, a seleção da terminologia mais frequente, a montagem do corpus e os demais procedimentos merecem mais atenção. A preocupação no tocante não só ao conteúdo a ser ensinado, mas também às palavras mais frequentes é uma questão fulcral.

Parece haver um certo distanciamento entre algumas áreas da linguística. Ou melhor, os linguistas de corpus raramente recorrem aos preceitos da linguística aplicada. Por sua vez, os estudiosos da linguística aplicada, em geral, pouco conhecimento possuem sobre a linguística de corpus. Uma interação da linguística aplicada e a linguística de corpus poderia revelar caminhos profícuos em um futuro próximo. O aporte de ambas as áreas tende a enriquecer os estudos linguísticos.

REFERÊNCIAS BIBLIOGRÁFICAS

BABINI, M.; MARRANGHELLO, N. (2007). *Introdução às redes neurais artificiais*. São Paulo: Cultura Acadêmica; São José do Rio Preto: Laboratório Editorial Ibilce.

BARROS, L.A. (2004). *Curso básico de terminologia*. São Paulo: Editora da Universidade de São Paulo.

BERBER SARDINHA, T. (2000). Linguística de corpus: histórico e problemática. *Delta*, v. 16, n. 2, p. 323-367.
_____. (2005). Como encontrar as palavras-chave mais importantes de um corpus com Wordsmith Tools. *Delta.* v. 2, n. 21, p. 237-250.

BIBER, D.; CONRAD, S. & REPPEN, R. (1998). *Corpus Linguistics*: Investigating language structure and use. Cambridge: CUP.

BIDERMAN, M. T. (2001). *Teoria linguística*: teoria lexical e computacional. 2 ed. São Paulo: Martins Fontes.

CABRÉ, M. T. (1993). *La terminologia*: teoria, metodologia, aplicaciones. Barcelona: Editorial Antartida/ Empuries.

_____. (2007). *Constituir um corpus de especialidad*: condiciones y posibilidades. Disponível em: <www.upf.edu/pdi/dtf/teresa.cabre/docums/ca07arra.pdf>. Acesso em: 14 set. 2008.

COBB, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology.* v. 11, n. 3, p. 38-63.
_____. (2008). *Web VP*. Disponível em: < http://www.lextutor.ca/vp/eng>. Acesso em: 02 nov. 2008.

COXHEAD, A.; HIRSH, D. (2007). A pilot science word list for EAP. *Revue française de linguistique appliqueé.* v. 7, n. 2, p. 65-78.

FERREIRA, A. B. H. (2004). *Novo dicionário Aurélio século XXI*: dicionário eletrônico versão 5.0. Curitiba: Positivo.

GÓMEZ, P. C. (2002). Do we need statistics when we have linguistics? *Delta.* v. 2, n. 18, p. 233-271.

NATION. I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: CUP.
_____. (2003). *Como estruturar o aprendizado de vocabulário*. São Paulo: SBS.

ORENHA, A. (2004). O uso de corpora e o ensino de ESP. *Estudos linguísticos***:** XXXIII, p. 1036-1041.

SCARAMUCCI, M. R. V. (1997). A competência lexical de alunos universitários aprendendo a ler em inglês como língua estrangeira. *Delta*. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44501997000200003> Acesso em: 05 abr. 2009.

SCOTT, M. (1999). *WordSmith Tools***:** version 3. Oxford: Oxford University Press.
_____. (2005). *Corpus Linguistics and ESP***:** is there a link? Disponível em <http://www.lexically.net/downloads/corpus_linguistics/CL_ESP_PUC_2005.ppt>. Acesso em: 26 ago. 2008.

SILVA, E.B. (2009). *Proposta de um dicionário eletrônico terminológico onomasiológico bilíngue inglês-português no domínio das Redes Neurais Artificiais*. Dissertação de Mestrado em Estudos Linguísticos, Instituto de Biociências, Letras e Ciências Exatas, Unesp, São José do Rio Preto.

SINCLAIR, J. (2004). Corpus and Text: Basic Principles. In: WYNNE, M. (ed.). *Developing linguistic corpora*: a guide to good practice. Disponível em: <http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm>. Acesso em: 14 set. 2008.

VIDAL, V; CABRÉ, M. T. (2005). *Estrategias para la ensenanza de combinaciones léxicas metaforicas en un curso de lenguas para fines específicos.* Disponível em: <www.upf.edu/pdi/dtf/teresa.cabre/docums/ca06aela.pdf>. Acesso em: 30 nov. 2007.

Recebido: 06/04/2009
Aceito: 18/03/2011