DOMÍNIOS DE LINGU@GEM
Revista Eletrônica de Linguística (http://www.seer.ufu.br/index.php/dominiosdelinguagem)
Volume 5, - n° 1 – 1° Semestre 2011 - ISSN 1980-5799

VocabProfile: uma ferramenta linguístico-estatística para a aula de língua inglesa

Eduardo Batista da Silva*

Resumo: Tanto o aporte da linguística computacional como as contribuições da linguística de *corpus* colaboram para tornar o ensino-aprendizagem de língua inglesa mais profícuo. O objetivo geral deste trabalho é apresentar o *Vocabprofile* (VP), versão 3, um programa para análise linguística que divide o texto em várias faixas de frequência lexical. Será apresentada também a *General Service List*, uma lista com as palavras mais comuns da língua inglesa, que faz parte do banco de dados do VP. A fim de exemplificar o funcionamento do *software*, utilizamos inicialmente um trecho da obra literária escrita em língua inglesa e de domínio público: *The Picture of Dorian Gray*. Na página de resultados, o texto é recortado em partes que são indicadas em quatro categorias: 1) *K1 words*: as primeiras 1000 palavras mais frequentes da língua inglesa, as palavras fundamentais, 2) *K2 words*: as próximas 1000 palavras mais frequentes da língua inglesa, 3) AWL: As palavras utilizadas em textos científicos de diversas áreas de especialidades e 4) *OFF-list*: As palavras "difíceis" que não se encontram em nenhuma das outras três listas anteriormente mencionadas. Com relação ao trecho selecionado para estudo, os cálculos estatísticos do VP apontam o seguinte perfil lexical: K1= 85,50%, K2= 6,11%, AWL= 0,76% e OFF= 7,63%. Efetuamos também uma comparação entre dois *corpora* de gêneros distintos sob os auspícios do VP.
Palavras-chave: *VocabProfile*; Linguística Computacional; Vocabulário; Língua Inglesa.

Abstract: Both the approach of computational linguistics and the contributions of corpus linguistics collaborate to make the teaching-learning process of the English language more profitable. The aim of this paper is to introduce *Vocabprofile* (VP), version 3, a software for linguistic analysis that divides a text into several frequency levels. A *General Service List*, a list with the most common words in English Language which take parto n VP database will be presented. In order to exemplify the software functioning, we used an excerpt of a literary work in English and of free domain: *The Picture of Dorian Gray*. In the output page, the text is broken up into parts that are sorted into four categories: 1) K1 words: the first 1000 most common words of the English language, the most important words, 2) K2 words: the next 1000 most common words of the English language, 3) AWL: the words used in scientific texts of several specialty areas and 4) OFF-list: the "difficult" words that are not found in any of the other previously mentioned lists. Regarding the excerpt selected for analysis, the VP statistic calculus show the following lexical profile: K1= 85,50%, K2= 6,11%, AWL= 0,76% e OFF= 7,63%. We also carried out a comparison with other corpora of distinct genres by resorting to VP.
Keywords: VocabProfile; Computational Linguistics; Vocabulary; English Language.

* Professor de Língua Inglesa da Universidade Estadual de Goiás, Quirinópolis – GO / Doutorando em Estudos Linguísticos pela Unesp/Ibilce, São José do Rio Preto – SP.

Introdução

Othero e Menuzzi (2005, p. 12) afirmam que a linguística computacional é a área da linguística que se ocupa do tratamento computacional da linguagem para diversas finalidades práticas. Costuma-se dividir a linguística computacional em duas subáreas: a linguística de *corpus* e o processamento de linguagem natural, também conhecido como PLN.

Graças aos avanços da informática, especialmente após os anos 1980, diversas áreas do conhecimento puderam observar e ao mesmo tempo manipular seus objetos de estudo sob a perspectiva da chamada linguística computacional, de fato, inovadora e empírica. A possibilidade de manipulação da língua pelo computador agiliza a realização de análises linguísticas variadas. Com o aporte da informática, atesta-se que a pesquisa de viés linguístico é incrementada sobremaneira, especialmente o ensino-aprendizagem de línguas estrangeiras. Assim, ao demonstrar como a língua se comporta em determinadas situações reais de uso, o pesquisador/professor encontra meios de comprovar ou refutar hipóteses.

Apesar de todas as vantagens que a linguística computacional proporciona atualmente,

> [...] se perguntarmos à maioria dos cidadãos se eles acham que a informática está presente nas áreas vistas como menos tecnológicas, como as da esfera das Ciências Humanas, e Letras, em particular, a maioria das respostas, com quase certeza, seria negativa." (BERBER SARDINHA, 2005, p. 8).

Deduz-se que ainda não está estabelecida na consciência do grande público a participação positiva dos recursos da informática nos estudos linguísticos. A noção que a maioria dos cidadãos compartilha é de que o universo das letras pode prescindir de tecnologia em seus mais diversos segmentos. É comum encontrar pesquisadores dedicados aos estudos linguísticos ou literários abstraindo-se da tecnologia disponível.

O ensino-aprendizagem de língua inglesa no suporte eletrônico torna a manipulação das informações mais ágeis e menos propensas a erros. Outrossim, a investigação e o acesso aos dados em uma base computacional são vantagens para os consulentes por duas razões: gerenciamento e compartilhamento de informações. Com relação ao ensino, Perrenoud (2000, p. 125) observa que a escola não pode ignorar o que se passa no mundo. Acrescenta, ainda, que as novas tecnologias da informação e

comunicação transformam espetacularmente não só nossas maneiras de comunicar, mas também de trabalhar, de decidir, de pensar.

O ensino-aprendizagem de língua inglesa pode ser enriquecido com o auxílio dos recursos computacionais, fazendo com que o trabalho de investigação linguística seja em larga escala abreviado sem que a qualidade da pesquisa seja afetada negativamente. O pesquisador não precisa mais ler ou conferir milhares de páginas em busca de determinadas palavras e/ou estruturas gramaticais, gastando dias ou meses nessa atividade. Com a intervenção da informática, todo o processo leva apenas alguns segundos.

O software VocabProfile

Nos últimos anos, diversos *softwares* vêm sendo desenvolvidos especialmente para a análise linguística. Neste contexto, o *software Vocabprofile* (VP), versão 3, pode auxiliar no estudo de inglês e merece, portanto, destaque por dinamizar a maneira de olhar textos, seja qual for o gênero estudado. Dessa forma, o VP viabiliza procedimentos para a análise qualitativa e quantitativa do léxico.

Desenvolvido em 2001 por Tom Cobb, professor da Universidade de Laval, no Québec, o VP sofreu desde então algumas modificações e está em sua terceira versão, otimizada para processar aproximadamente 1500 palavras por segundo. Foi inspirado no *software* de Heatley e Nation (1994), chamado *Range.* Diferentemente do *Range*, o VP só pode ser utilizado na plataforma *online*. Por essa razão, sempre que não for possível usar o VP pela internet, seja por falha de conexão ou por sobrecarga no sistema, o *software Range* – apesar de sua simplicidade – pode ser de grande valia, já que pode ser armazenado no disco rígido de qualquer computador pessoal.

O VP está hospedado em um portal voltado para o aprendizado de línguas estrangeiras, principalmente a língua inglesa e, em menor escala, a língua francesa. O portal é chamado *Compleat Lexical Tutor* (www.lextutor.ca) e é mantido por Tom Cobb. O interesse gerado pelos recursos linguísticos do portal pode ser atestado pelo aumento no número de páginas carregadas ao longo dos anos. Em 2005, foram carregadas 213.528 páginas dentro do *Compleat Lexical Tutor*. Em contraste, no ano de 2010, o total de páginas carregadas alcançou o montante de 2.751.075. Sem dúvida, tal aumento expressivo revela que mais e mais pessoas de diversas partes do mundo estão

realizando pesquisas linguísticas no portal, sendo os motivos diversos: pesquisas relacionadas ao vocabulário, questões acadêmicas ou simples curiosidade. Alternativamente, existe a possibilidade de o VP ser usado por estudantes para checar sua amplitude lexical e a densidade de produção de vocabulário (COBB, s/d).

Se, por um lado, os números apresentados acima impressionam, por outro, é importante destacar que a preocupação com os estudos que envolvem o léxico não são exclusividade da era da informática. Já na primeira metade do século XX, o vocabulário essencial ganha vulto nos estudos linguísticos. Em 1921, Thorndike publica *Teacher´s Word Book*. Em 1930, Ogden formula um léxico básico de 850 palavras com um princípio: ter a maior abrangência possível usando a menor quantidade possível de palavras com sentido geral. Em 1932, Thorndike publica *A Teacher´s Word Book of 20,000 Words*. Thorndike e Lorge elaboraram, em 1944, *The Teacher´s Word Book of 30,000 Words*. A proposta do livro era auxiliar no ensino de inglês para alunos falantes de outras línguas. Ao consultar o livro em questão, o professor teria condições de selecionar o vocabulário ideal para os aprendizes.

No que concerne estudos de levantamento de vocabulário, interessa especialmente neste trabalho ressaltar a importância de um livro lançado no ano de 1953: *A General Service List of English Words*. Seu autor, Michael West, foi influenciado, em grande parte, pela obra anterior de Thorndike e Lorge e pelas listas de palavras elaboradas em anos anteriores. O princípio norteador destes trabalhos era a frequência das palavras. Consequentemente, as palavras mais frequentes mereciam mais atenção na sala de aula de língua inglesa. A lista de palavras presentes em *A General Service List of English Words* também é conhecida como GSL. Tal lista é a base lexical de referência que o VP utiliza para fornecer os resultados de suas análises.

Apresentando o programa

Ainda que o *layout* do *software* contenha uma alta densidade de informações, sua utilização é relativamente simples. Assim que a página *VocabProfile Home* (http://www.lextutor.ca/vp/) é carregada, pode-se identificar as quatro versões independentes disponíveis para se pesquisar textos em língua inglesa, como pode ser comprovado pela Figura 1. Encontram-se na *homepage* as três versões experimentais, a saber: BNC-20, BNL e *Kids*, além da versão denominada *Classic*. Cabe ao pesquisador

decidir qual delas se adequa mais ao propósito da investigação a ser realizada, uma vez que cada versão possui suas especificidades.

A primeira versão aqui elencada, a BNC-20, utiliza parte das palavras do *British National Corpus*, cujo banco de dados ultrapassa a marca de 100 milhões de palavras. As pesquisas executadas com o BNC retornam 20 faixas de frequência, cada uma com 1000 famílias de palavras (o conceito de família será explicado mais adiante). Se determinada palavra não pertencer a nenhuma das faixas, ela será aninhada em uma lista de exclusão, chamada *OFF-list*. Como cada faixa de frequência recebe uma cor, de tonalidades próximas, às vezes, torna-se um pouco complicada a interpretação dos resultados. Ainda assim, um exercício curioso é a constatação do papel seminal que as primeiras faixas de frequência desempenham nos textos em geral, especialmente as duas primeiras, que possuem um índice de abrangência de aproximadamente 90%.

O BNL (*Bare Naked Lexis*) é uma versão experimental, desenvolvida por dois pesquisadores turcos, que divide o texto em seis faixas de frequência, sendo a BNL-0 a primeira e a BNL-6, a última. Cada faixa possui aproximadamente 450 palavras. As palavras que não se enquadram em nenhuma das seis listas são deslocadas para a *OFF-list*. O BNL surgiu para auxiliar estudantes de inglês na leitura de textos acadêmicos e possui 2.709 palavras lematizadas em seu banco de dados. Em média, 90% dos textos acadêmicos são constituídos dessas palavras, selecionadas para o estudante universitário que não tem o inglês como primeira língua.

A versão *Kids* propõe-se a identificar o incremento lexical nas crianças. O *VP-Kids* fornece resultados usando 10 faixas de frequência, cada uma com 250 palavras lematizadas, fruto de diversos estudos empíricos sobre produção oral infantil. Existe também, na página de resultados, uma lista de exclusão "conhecida" (com as palavras que as crianças podem reconhecer, mas que geralmente não utilizam) e uma lista de exclusão "desconhecida" (com palavras mais formais e nomes próprios).

Demonstraremos a seguir, algumas das principais funcionalidades da versão *Classic*, que é a mais popular e simples, sendo a versão ideal para um primeiro contato com o *software*. O *link* para a versão *Classic* encontra-se logo na primeira linha da página: Classic VP English v.3.

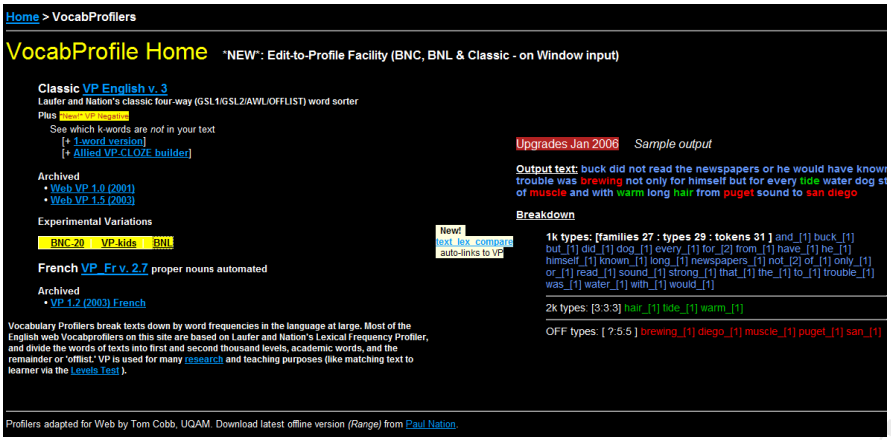

Figura 1: a página inicial do VP

Na parte inferior, à esquerda, pode-se visualizar um *link* chamado *Levels Test*, cujo objetivo é apresentar testes lexicais variados. O recurso em questão pode ser explorado para fins didáticos bastante úteis na avaliação do nível de conhecimento lexical de determinado estudante ou grupo. Os resultados de alguns testes podem ser obtidos assim que as respostas são enviadas, viabilizando, por exemplo, a aplicação dos testes em um laboratório de informática.

Do lado direito, existe uma pequena demonstração de como o VP fornece sua análise linguístico-estatística. Outros detalhes sobre esses resultados serão discutidos mais adiante.

Inserindo um texto

A Figura 2 revela que existem duas maneiras distintas de inserção de texto, o método A (com fundo de tela preto) e o método B (com fundo de tela na cor vermelha). No primeiro método, o usuário digita ou cola um texto na área mais clara da página, que é uma caixa de texto. Antes disso, naturalmente, é necessário deletar as instruções ali contidas. Isso pode ser feito rapidamente com um comando simples, Ctrl + A, que seleciona todo o conteúdo e, posteriormente, pressionando a tecla “del” ou “backspace”. Após o comando Ctrl + V, o texto é colado na tela e o próximo passo é clicar no botão SUBMIT, localizado na parte central, à direita da tela. É recomendado inserir textos constituídos de menos de 2000 palavras ou até 30000 caracteres.

No segundo método (com fundo de tela na cor vermelha), indicado para textos grandes, o arquivo em formato texto simples, sem formatação, deve ser selecionado no disco rígido do computador e enviado pelo botão *Submit File*. Ao processar arquivos

textuais mais robustos, próximos de 1,5 Mb, o VP começa a apresentar limitações na velocidade de processamento e tende a retornar páginas de erro.

Figura 2: a versão *Classic* do VP

Conceitos importantes

Ao inserir um texto na janela de consulta, como demonstrado na Figura 2, o VP executa a análise cruzada com seu banco de dados, onde está instalada a lista de palavras GSL, retornando uma página de resultados bastante completa e com diversos dados. O resultado final obtido pelo VP pode ser editado, exportado, disponibilizado em rede intranet, impresso ou publicado em meio virtual.

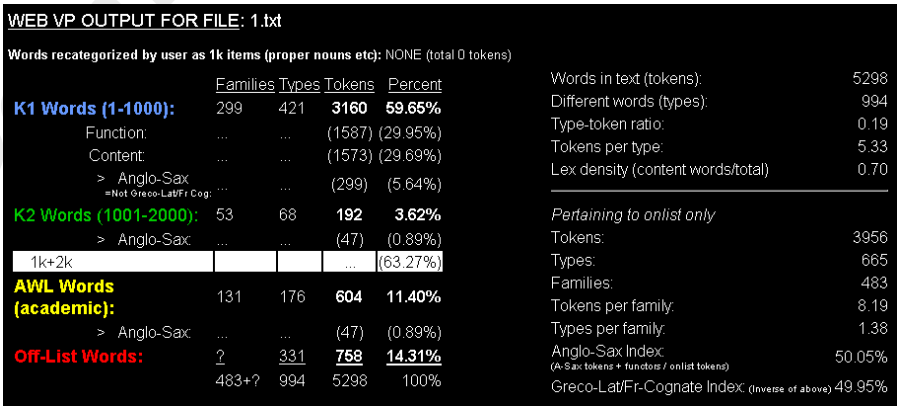

Figura 3: apresentação dos dados linguístico-estatísticos

A fim de interpretar os dados linguístico-estatísticos apresentados, faz-se necessária uma apresentação de alguns termos fundamentais, como *families*, *types*, *tokens*, *K1 words*, *K2 words*, *AWL* e *OFF-List*.

- ***Families*:** as famílias são as palavras que compartilham da mesma raiz lexical. As formas derivadas acabam sendo subordinadas à forma base. Esse agrupamento também é conhecido como lematização. A fim de facilitar a identificação dos lemas, eles serão escritos em versalete. Por exemplo, na família do verbete (*headword*) *ACCURATE* estão agrupados *accuracy*, *accurately*, *inaccuracy*, *inaccuracies* e *inaccurate*. Do mesmo modo, as formas conjugadas no *simple present* (*am, are, is*), no *present continuous* (*being*) e no *simple past* (was, were) acabam sendo subordinadas ao verbete *BE*. Neste caso, tem-se 1 família, 6 *types* e 6 *tokens*.
- **Types:** Em língua portuguesa, é possível traduzir *types* como formas ou vocábulos (BERBER SARDINHA, 2004). No entanto, utilizaremos a palavra em língua inglesa, *type*. Trata-se de cada palavra considerada isoladamente, sem repetição, no texto. Por exemplo, a frase "I am who I am", possui 3 types (I, am, who) e 5 tokens (I, am, who, I, am)
- **Tokens:** Também traduzido como itens ou ocorrências (BERBER SARDINHA, 2004). Todas as ocorrências de todas as palavras presentes no texto são contabilizadas. Na maioria das vezes, os textos possuem um número maior de *tokens* que de *types*.
- ***K1 words*:** Nessa faixa, encontram-se as palavras fundamentais da língua inglesa. São as primeiras 1000 palavras mais frequentes da GSL. Na página de resultados, o número que aparece entre colchetes ao lado da palavra indica a quantidade de vezes que a mesma ocorreu no texto pesquisado. Cumpre destacar que a categoria gramatical mais comum nessa faixa é a dos substantivos. Em segundo e terceiro lugar, aparecem os verbos e os adjetivos, respectivamente. As demais categorias gramaticais aparecem em menor número. As palavras pertencentes a esta categoria são destacadas pelo VP na cor azul. Segue uma amostra das palavras mais frequentes da GSL iniciadas pela letra "n": *NAME, NARROW, NATION, NATIVE, NATURE, NEAR, NECESSARY, NECK, NEED, NEIGHBOUR, NEITHER, NEVER, NEW, NEXT, NIGHT, NO, NONE, NOR, NORTH, NOT, NOTE, NOTHING,*

NOTICE, *NOW*, *NOWHERE* e *NUMBER*. Uma vez que as palavras não recebem etiquetagem morfossintática no VP, não é possível saber se *NEED*, por exemplo, é referente ao substantivo ou ao verbo. No entanto, West (1953) indica que *NEED*, na GSL, ocorre 63% das vezes como verbo e 26% como substantivo. Os adjetivos *needful*, *needless* e *needy* são responsáveis pelas demais ocorrências, 11%. O VP apenas destaca a ocorrência de uma palavra, mas não é capaz de fornecer dados mais precisos de uso ou informações de ordem léxico-gramatical.

- ***K2 words*:** A segunda faixa de palavras mais frequentes inclui as palavras da posição 1001 à 2000 da GSL. O VP as identifica pela cor verde. Curiosamente, o conhecimento desta faixa de palavras acrescenta pouco entendimento na leitura de um texto, quando comparado ao índice de abrangência da lista K1. Nation (2003) indica que essas 1000 palavras elevam em torno de 5% o conhecimento das palavras de um texto, ao passo que o conhecimento das primeiras 1000 palavras, da lista K1, conseguem abarcar cerca de 80% de um texto. Na sequência, uma amostra das palavras iniciadas pela letra "h", pertencentes à lista K2: *HAIR*, *HAMMER*, *HANDKERCHIEF*, *HARBOUR*, *HARM*, *HARVEST*, *HASTE*, *HAT*, *HATE*, *HAY*, *HEAL*, *HEAP*, *HEART*, *HEIGHT*, *HESITATE*, *HINDER*, *HIRE*, *HIT*, *HOLE*, *HOLIDAY*, *HOLLOW*, *HOLY*, *HONEST*, *HOOK*, *HORIZON*, *HOSPITAL*, *HOST*, *HOTEL*, *HUMBLE*, *HUNGER*, *HUNT*, *HURRY*, *HURT e HUT*. Com relação ao verbete *HUMBLE*, West (1953, p. 241) indica dois sentidos e traz dois exemplos de uso: 1) *modest – in a humble voice*; *My humble opinion* e 2) *not distinguished – Of humble birth*; *a humble home*. O primeiro item ocorre em 33% dos casos, ao passo que o segundo item em 59%. O advérbio *humbly* responde pelas outras ocorrências. Tais informações poderiam fazer parte do VP, proporcionando mais perspectivas de investigação e estudo da língua inglesa.
- **AWL:** A *Academic Word List* (AWL), a Lista de Palavras Acadêmicas é constituída de 570 famílias de palavras. É a única lista que, originalmente, não faz parte da GSL. Ela foi desenvolvida por Coxhead (2000) e incluída no banco de dados do VP por contemplar as palavras mais comuns dos textos técnico-científicos. Nation (2003, p. 61) explica que AWL foi elaborada a partir de um *corpus* com 3.600.000 palavras, contendo textos acadêmicos de quatro seções principais: Artes, Comércio, Ciência e Direito. Cada seção foi subdividida em

sete subseções. Para constarem na AWL, as famílias de palavras deveriam ocorrer nas quatro seções principais e, em pelo menos, 15 das 28 subseções. Chegou-se a um número final de 10 listas de palavras acadêmicas. Segundo Nation (op.cit.), essa lista engloba 8,5%-10% das palavras que ocorrem em textos acadêmicos. Pouco mais de 1% das palavras utilizadas em romances e na conversação faz parte da AWL. As palavras abaixo são aquelas que mais aparecem nos textos acadêmicos. Todas pertencem à lista 1, aquela que inclui as palavras mais comuns nesse gênero. Como é possível perceber, a lista já se encontra lematizada, mostrando apenas os verbetes (*headwords*): *ANALYSE*, *APPROACH*, *AREA*, *ASSESS*, *ASSUME*, *AUTHORITY*, *AVAILABLE*, *BENEFIT*, *CONCEPT*, *CONSIST*, *CONSTITUTE*, *CONTEXT*, *CONTRACT*, *CREATE*, *DATA*, *DEFINE*, *DERIVE*, *DISTRIBUTE*, *ECONOMY*, *ENVIRONMENT*, *ESTABLISH*, *ESTIMATE*, *EVIDENT*, *EXPORT*, *FACTOR*, *FINANCE*, *FORMULA*, *FUNCTION*, *IDENTIFY*, *INCOME*, *INDICATE*, *INDIVIDUAL*, *INTERPRET*, *INVOLVE*, *ISSUE*, *LABOUR*, *LEGAL*, *LEGISLATE*, *MAJOR*, *METHOD*, *OCCUR*, *PERCENT*, *PERIOD*, *POLICY*, *PRINCIPLE*, *PROCEED*, *PROCESS*, *REQUIRE*, *RESEARCH*, *RESPOND*, *ROLE*, *SECTION*, *SECTOR*, *SIGNIFICANT*, *SIMILAR*, *SOURCE*, *SPECIFIC*, *STRUCTURE*, *THEORY* e *VARY*.

Estas são as palavras acadêmicas da lista 10: *ADJACENT*, *ALBEIT*, *ASSEMBLE*, *COLLAPSE*, *COLLEAGUE*, *COMPILE*, *CONCEIVE*, *CONVINCE*, *DEPRESS*, *ENCOUNTER*, *ENORMOUS*, *FORTHCOMING*, *INCLINE*, *INTEGRITY*, *INTRINSIC*, *INVOKE*, *LEVY*, *LIKEWISE*, *NONETHELESS*, *NOTWITHSTANDING*, *ODD*, *ONGOING*, *PANEL*, *PERSIST*, *POSE*, *RELUCTANCE*, *SO-CALLED*, *STRAIGHTFORWARD*, *UNDERGO* e *WHEREBY*.

Para o falante de língua portuguesa, diversas palavras da AWL não apresentam dificuldade de compreensão porque são cognatas, de origem latina, fato esse que auxilia no estudo do vocabulário acadêmico. Nation (2003) aponta que as palavras acadêmicas devem ser idealmente estudadas após o domínio das primeiras 2000 palavras.

- ***OFF-list*:** As palavras *OFF-list* são aquelas que não se encontram em nenhuma das listas mencionadas anteriormente. As palavras desta categoria ganham a cor vermelha. Aqui, pode-se encontrar todos os nomes próprios, palavras que não foram digitadas corretamente, palavras estrangeiras, palavras mais formais e termos de áreas de especialidade.

Aplicações do VP no ensino de língua inglesa

O trecho a seguir foi retirado do livro *The Picture of Dorian Gray*, são os dois últimos parágrafos do livro de Oscar Wilde. A visualização das palavras fornecida pelo VP torna viável a manipulação do texto para que os aprendizes sintam-se mais confortáveis com as palavras ou para que o pesquisador ateste alguma produção linguística.

> After about a quarter of an hour, he got the coachman and one of the footmen and crept upstairs. They knocked, but there was no reply. They called out. Everything was still. Finally, after vainly trying to force the door, they got on the roof and dropped down on to the balcony. The windows yielded easily--their bolts were old.
> When they entered, they found hanging upon the wall a splendid portrait of their master as they had last seen him, in all the wonder of his exquisite youth and beauty. Lying on the floor was a dead man, in evening dress, with a knife in his heart. He was withered, wrinkled, and loathsome of visage. It was not till they had examined the rings that they recognized who it was.

Figura 4: Relação de types identificados no texto

Na Figura 4, pode-se visualizar a classificação das palavras de acordo com sua frequência na GSL e na AWL. As palavras são elencadas de acordo com a ordem alfabética e não pela frequência no texto. Por exemplo, o artigo definido *THE* foi usado 10 vezes e aparece na antepenúltima linha de resultado da faixa K1.

Muitos professores e desenvolvedores de cursos usam o *VocabProfile* para modificar o perfil lexical dos textos instrucionais que serão praticados com os alunos (COBB, 2010). Com o VP, o professor de língua inglesa ou qualquer outro usuário do programa é capaz de avaliar se determinado texto é, de fato, adequado para um aluno ou uma turma. O nível de dificuldade lexical pode ser, até certo ponto,

mensurado. Partindo desse princípio, uma turma de nível avançado poderia praticar as palavras mais complexas ou menos frequentes do trecho. É possível destacar os seis substantivos (*balcony*, *bolt*, *coachman*, *footman*, *portrait* e *visage*) e os dois verbos (*wither* e *wrinkle*) da lista de exclusão. Diferentes tipos de exercícios podem ser preparados a partir das palavras selecionadas. Oportunamente, West (1953, p. ix) atesta que a frequência não é o único quesito a ser considerado na seleção de palavras no ensino de inglês. Outros fatores devem ser considerados, bem como 1) facilidade ou dificuldade de aprendizado, 2) necessidade, 3) abrangência, 4) estilo e 5) emoção.

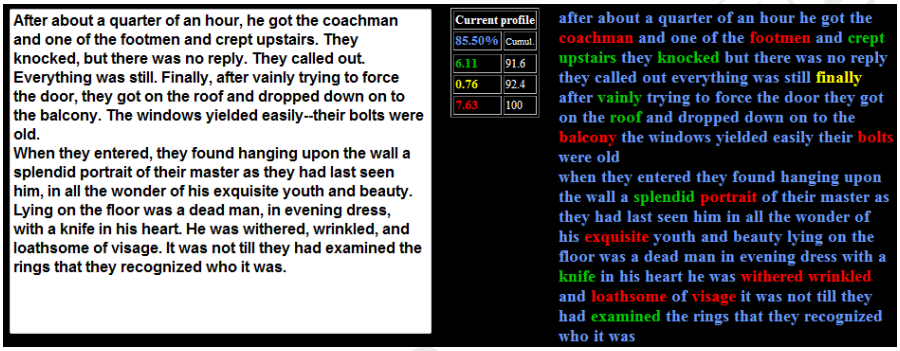

Figura 5: visualização e análise qualitativa do texto

Em outro ponto da página de resultados, como mostra a Figura 5, é possível cotejar o texto original e o texto analisado. As dimensões da tela de fundo branco, onde o texto original é apresentado, podem ser alteradas, em função da necessidade do consulente. Com relação ao trecho selecionado para estudo, os cálculos estatísticos do VP apontam o seguinte perfil lexical: K1= 85,50%, K2= 6,11%, AWL= 0,76% e OFF= 7,63%. O domínio das primeiras 2000 palavras, das listas K1 e K2, garantiria o reconhecimento de 91,61% do vocabulário do trecho selecionado para estudo. Tais achados servem para guiar inúmeros tipos de pesquisas linguísticas voltadas à prática de vocabulário.

Com o intuito de demonstrar uma consulta mais elaborada realizada com o VP, recorremos a dois *corpora*, previamente trabalhados no âmbito do Grupo de Estudos em Análise Linguística e Linguística de *Corpus* – GEALLC – da Universidade Estadual de Goiás. O objetivo aqui é trabalhar com uma quantidade de palavras significativamente maior e comprovar algumas hipóteses relacionadas às palavras contidas nos textos técnico-científicos e nos textos musicais.

Tabela 1: *corpora* com respectivos perfis lexicais

	K1	K2	AWL	OFF
Textos científicos	59,65%	3,62%	11,40%	14,31%
Textos musicais	84,87%	5,97%	0,49%	8,65%

O *corpus* técnico-científico caracteriza-se pela presença de palavras mais especializadas. Existe uma preocupação com a formalidade e com as estruturas gramaticais do texto. O *corpus* técnico-científico contém textos que abordam as Redes Neurais Artificiais, uma subárea da Inteligência Artificial.

O outro *corpus*, de letras de músicas, serve para representar a língua oral, com suas características mais informais e com construções mais simples. O *corpus* de música possui 782 letras de seis artistas: Nirvana, Metallica, Michael Jackson, Iron Maiden, Beatles e Queen. No tocante às letras de músicas, Silva (2010), utilizando o VP, mostra que quase 85% das palavras das letras em questão possuem um vocabulário considerado básico, ou seja, enquadram-se na faixa das 1000 palavras mais frequentes da língua inglesa, a lista K1. Por outro lado, as palavras difíceis, constituem menos de 10% do *corpus*, sendo principalmente nomes próprios e palavras que estão fora do escopo das demais listas produzidas pelo VP. Vale destacar que o conhecimento das palavras mais frequentes, o "núcleo duro", pode servir de base e ser explorado também em outras atividades relacionadas à língua, a saber: a leitura, a oralidade, a compreensão auditiva e a redação. Assim sendo, uma abordagem especificamente direcionada à apresentação, treinamento e fixação de vocabulário contido em letras de música tende a fornecer resultados satisfatórios aos aprendizes.

Com relação às 1000 palavras mais comuns (K1), é possível perceber que os textos musicais em questão possuem uma concentração de palavras fundamentais 42% maior, quando comparados aos textos técnico-científicos das Redes Neurais Artificiais. Isso se deve a propriedades do gênero linguístico, que, neste caso, exige uma comunicação voltada a um público já especializado, fazendo com que haja uma preocupação relacionada ao nível de vocabulário empregado, além da terminologia da área.

A segunda faixa de palavras mais comuns (K2) é pouco percebida nos *corpora* analisados. Porém, seu uso é ligeiramente superior nos textos com as letras de músicas.

Salta aos olhos a discrepância entre a porcentagem de palavras acadêmicas (AWL) que foi identificada nos dois *corpora* de estudo. Nos textos técnico-científicos, as palavras acadêmicas aparecem 23 vezes mais que nos textos com letras de músicas. Devido à própria natureza das letras de músicas, de simular uma linguagem menos formal, mais relaxada, não são percebidas as palavras acadêmicas de forma peremptória.

Finalmente, percebe-se que os textos técnico-científicos estudados apresentam uma alta porcentagem de palavras "desconhecidas", o que é indicado na *OFF-list*. Conforme demonstrado há pouco, as palavras mais raras, dentre outras, foram alocadas nessa categoria. A porcentagem de palavras de exclusão nas letras de músicas, por sua vez, é mais reduzida.

Considerações Finais

O cotidiano da sala de aula de língua inglesa permanece, salvo raras exceções, enraizado basicamente na mesma tradição do século passado: lousa/giz, na qual o professor fala e o aluno copia. Infelizmente, poucos professores possuem conhecimentos relacionados a *softwares* como o VP. O presente trabalho teve como objetivo geral apresentar as funcionalidades do VP que podem facilmente ser utilizadas por qualquer indivíduo envolvido com o ensino-aprendizagem de vocabulário em língua inglesa.

Apesar de todos os avanços da tecnologia em diversas áreas, o tratamento computacional da língua ainda permanece pouco difundido. Via de regra, alunos e professores do curso de Letras tendem a evitar recursos tecnológicos em seus trabalhos, seja por medo ou desconhecimento.

No entanto, as pesquisas realizadas nas últimas décadas mostram que uma maior proximidade entre áreas distintas como a Linguística Computacional, a Linguística de *Corpus* e a Linguística Aplicada pode enriquecer as investigações tanto de base lexical quanto estrutural.

No que se refere à utilização de novas tecnologias na educação, os resultados indicam que elas são capazes de atrair a atenção dos alunos, quebrando paradigmas educacionais e maximizando o contato com a língua inglesa, principalmente com o

vocabulário. Depreende-se disso que quanto mais familiaridade o professor possui com os avanços computacionais, maiores podem ser as taxas de êxito, nomeadamente: aprendizagem, motivação e interação.

Com o VP, o pesquisador ou o professor pode selecionar o vocabulário com o qual deseja trabalhar. Caso os alunos não dominem o vocabulário básico da língua inglesa, é possível recorrer às palavras presentes em um texto específico, ou mesmo em um *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.

Vale ressaltar que a ideia não é simplesmente tomar uma posição de espectador e deixar o VP liderar o processo educacional. Pelo contrário, as revelações fornecidas pelo *software* servem de subsídio para outras atividades didáticas. Cabe ao professor ou ao pesquisador a elaboração de estratégias que propiciarão entendimento e fixação de vocabulário. Caso as listas de frequência ou mesmo os resultados linguístico-estatísticos fossem simplesmente entregues nas mãos dos alunos, pouca ou nenhuma contribuição isso geraria. O trabalho manual com essas listas acaba sendo enfadonho por um simples motivo: não é atraente nem prático.

Graças ao tratamento computadorizado do léxico, uma maior e melhor manipulação das palavras presentes em um texto é viabilizada. Pode-se recorrer a várias listas simultaneamente, obtendo resultados fiáveis para a prática de vocabulário. As informações deste trabalho servem para guiar inúmeros tipos de pesquisas linguísticas, desde a seleção do nível de dificuldade lexical de um texto a ser utilizado com aprendizes até a análise de redações escritas em língua inglesa por estudantes brasileiros.

O vocabulário de uma língua estrangeira constitui importante material tanto de recepção quanto de produção linguística. Sem o domínio das palavras mais importantes do inglês, por exemplo, o entendimento de uma mensagem é prejudicado. Consequentemente, torna-se relevante travar contato com a língua inglesa por meio de *softwares* que forneçam informações linguístico-computacionais. A investigação linguística, especialmente aquela relacionada à língua inglesa, pode ser enriquecida com a utilização do VP.

Referências

BERBER SARDINHA, T. **A língua portuguesa no computador**. Campinas: Mercado de Letras, 2005. 296 p.

_____. **Linguística de corpus**. Barueri: Manole, 2004. 410 p.

COBB, T. **Some Research Uses of VocabProfile (VP)**. Disponível em: <http://www.lextutor.ca/vp/research.html>. Acesso em: 21 fev. 2011.

_____. Learning about Language and Learners from Computer Programs. **Reading in a Foreign Language**, v. 22, n. 1, p. 181–200, 2010.

COXHEAD, A. A New Academic Word List. **TESOL Quarterly**, v. 34, n. 2, p. 213-238, 2000.

HEATLEY, A.; NATION, P. **Range**: A Program for the Analysis of Vocabulary in Texts. 2004. Disponível em: <http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx>. Acesso em: 21 fev. 2011.

NATION, P. **Como estruturar o aprendizado de vocabulário**. Tradução de Cristiane Arruda. São Paulo: Special Book Services Livraria, 2003. 121 p.

OTHERO, G. A.; MENUZZI, S. M. **Linguística computacional**: teoria e prática. São Paulo: Parábola Editorial, 2005. 128 p.

PERRENOUD, P. **Dez novas competências para ensinar**. Tradução de Patrícia Chittoni Ramos. Porto Alegre: Artmed, 2000. 192 p.

SILVA, E. B. **Letras de música em língua inglesa:** um corpus a ser explorado. Comunicação apresentada à Semana de Letras da Universidade Federal de Uberlândia. Uberlândia, 2010. Não publicado.

WEST, M. **A General Service List of English Words**. London: Longmans and Green, 1953. 588 p.