Dados lexicométricos das provas do ENEM

Janaína Soares Martins[1] e José Pereira da Silva Júnior[2]

[1]Instituto Federal do Sul de Minas Gerais – Campus Machado, Machado, MG, janaina.smartins@hotmail.com [2]Instituto Federal do Sul de Minas Gerais – Campus Machado, Machado, MG, jpereira@mch.ifsuldeminas.edu.br

Introdução

Esta pesquisa buscou realizar o levantamento lexicométrico das provas aplicadas no Exame nacional do Ensino Médio (Enem), no período de 1998 a 2010, de maneira a fornecer um mapeamento dos termos com maior recorrência e suas correlações. Isso foi possível primeiramente com o uso de ferramentas de análise lexical que permitiram indexar os conceitos e termos recorrentes no corpus de análise, formado pelas questões do Enem. Essas informações poderão orientar não só os estudos de quem se prepara para o Enem, mas também os docentes em sua pratica pedagógica.

O Enem tornou-se um importante instrumento de avaliação do aprendizado de nível médio, sendo amplamente utilizado pelas instituições de ensino superior para validar o acesso de alunos aos seus cursos. Por isso tem crescido o interesse por esse exame, sendo oferecidos diversos cursos de preparação para os alunos interessados em atingir melhores notas. Atualmente, o IFSULDEMINAS oferece boa parte de suas vagas (70%) em cursos superiores de graduação vinculadas ao sistema SISU, que, por sua vez, se vale dos resultados do Enem.

Nesse contexto, a análise das provas do Enem facilita o estudo dos avaliados, que podem concentrar seus esforços nas temáticas e metodologias empregadas nas provas. Além disso, a partir dessa análise é possível reavaliar a prática mesma das escolas de nível médio de forma a questionar os conteúdos que trabalha ou, pela via oposta, aqueles que são destacados nas provas do Enem.

Instrumentalizar essa análise será de grande valia por essas razões. Entretanto o desafio que se coloca é sistematizar uma enorme quantidade de questões que constituem o acervo de anos de provas aplicadas. O esforço nessa análise pode ser, entretanto, bastante minimizado quando se utiliza a computação como ferramenta na leitura dessas questões.

A lexicometria é a medida do vocabulário de um texto ou conjunto de textos sob diversas perspectivas, utilizando-se de ferramentas de mineração de dados que podem facilitar sobremaneira essa tarefa, reduzindo bastante o tempo de análise de um corpus. A utilização da

computação nesse caso é fundamental, uma vez que a atividade para o ser humano seria demasiadamente dispendiosa inviabilizando sua realização.

Existem atualmente diversos softwares que realizam tarefas lexicométricas. O utilizado na pesquisa é o VBPRO, desenvolvido por MILLER (1995). É um conjunto de ferramentas *freeware* para utilização em pesquisas acadêmicas. Funciona em DOS, tornando-se extremamente rápido na sua operação.

A leitura artificial realizada por esse tipo de software é produzida através de algoritmos que fazem a contagem das palavras (strings) de um texto, produzindo informação a respeito de sua frequência e ranqueamento. Pode-se assim, utilizando-se o software adequado, descobrir informações que antes se encontrariam dispersas e seriam dificilmente percebidas.

Essa técnica empregada em levantamento de dados em grandes conjuntos de textos, os corpora, é utilizada com frequência em grandes instituições facilitando as tomadas de decisão de seus gestores. Aplicada às provas do Enem, podem ser descobertas informações a respeito de temáticas mais frequentes e suas correlações, facilitando a preparação dos avaliados.

Material e Métodos

A metodologia de pesquisa aplicada na extração de informações de um conjunto de textos consiste primeiramente na preparação desses textos para a leitura automática. Como o software utilizado funciona com a codificação ASCII *Standard* de caracteres, foi necessário preparar os documentos. Primeiramente, eles são convertidos do formato pdf para o txt, utilizando-se o Adobe Acrobat. Nessa conversão, entretanto, é produzida grande quantidade de caracteres não reconhecidos pelo VBPRO, o que interfere na análise. Para contornar essa dificuldade é necessário o ajustamento do formato de arquivo txt a ser lido pelo software.

Nessa etapa, é feita a seleção dos enunciados das questões eliminando as alternativas ou textos de apoio que poderiam distorcer a análise, por conterem vocabulário muitas vezes repetido ou fora do propósito da análise, que é saber os principais conteúdos referidos na prova.

A essa formatação do arquivo de texto segue-se a análise realizada pelo programa. Numa primeira etapa o arquivo de texto é formatado, gerando uma extensão frm. Em seguida, este arquivo frm passa por um processo de alfabetização em que são listadas todas as palavras do texto e sua frequência absoluta e relativa. Numa terceira etapa, esse resultado é ranqueado, produzindo uma listagem das palavras mais frequentes às menos frequentes.

A etapa seguinte da análise consistiu na retirada das palavras (*stopwords*) que interferem na análise do conteúdo como artigos, preposições, conjunções, verbos, adjetivos, etc. Devido ao alcance da pesquisa, não foram tratadas ambiguidades semânticas que demandariam uma análise particularizada. Essa indecisão, por exemplo, ocorre com palavras que podem pertencer a mais de uma classe gramatical, o que só pode ser resolvido na análise do contexto da frase. Para diminuir possíveis distorções, optou-se por analisar exclusivamente substantivos ou palavras que assumem formas idênticas a substantivos.

Resultados e Discussão

Esses substantivos oferecem uma visão dos conteúdos abordados na prova remetendo a três classes distintas:

1) palavras de conteúdo específico, particulares às áreas de conhecimento;

2) palavras de conteúdo geral, como operadores dos processos mentais (ex: referência, crescimento, diferença, etc), presentes em todas as áreas de saber;

3) palavras metatextuais, que se referem à leitura da prova, como texto, figura, tabela, etc.

Esses resultados permitiram uma visão dos principais temas abordados na prova (Tabela 1):

Tabela 1. Substantivos (categoria 1) mais frequentes, de maneira absoluta e relativa

Palavra	Frequência absoluta	Frequência relativa
energia	172	0.260
produção	76	0.115
região	72	0.109
população	65	0.098
vida	57	0.086
consumo	53	0.080
cidade	41	0.062
ambiente	31	0.047
gás	32	0.048
terra	29	0.044
massa	28	0.042
trabalho	28	0.042
temperatura	27	0.041
lixo	25	0.038
concentração	23	0.035

petróleo	23	0.035
solo	23	0.035
pH	23	0.035
luz	21	0.032
renda	21	0.032

Estes termos foram submetidos a uma segunda análise que indicou o campo lexical de cada um. Para fins de delimitação, foram analisados os campos lexicais das 20 principais ocorrências. Essa análise mostrou, por exemplo, qual o campo lexical da palavra energia, ou seja, na ocorrência desse termo quais outras são mais frequentes. Esse resultado foi alcançado com o agrupamento de todas as questões onde a palavra ocorreu produzindo um novo ranqueamento pelo mesmo processo referido acima.

Por exemplo, as questões que continham a palavra energia foram separadas em um documento, que após ser formatado gerando a extensão frm, passa por um processo de alfabetização em que são listadas todas as palavras do texto e sua frequência absoluta e relativa. Esse processo foi o mesmo utilizado anteriormente para gerar a lista de palavras mais relevantes.

Dos novos resultados obtidos do foi possível considerar uma nova lista de palavras. Observe a lista a segui com as palavras mais frequentes no contexto da palavra energia:

1) elétrica;
2) consumo;
3) produção;
4) processo;
5) eletricidade;
6) eficiência;
7) gás;
8) aquecimento;
9) combustíveis;
10) residência.

Com base nesses dados, podemos concluir que no campo lexical da palavra energia, a palavra elétrica é a palavra com maior relevância, ou seja, a que aparece com mais frequência nesse contexto.

O mesmo foi feito com as demais palavras (o processo foi realizado com as 20 palavras de maior relevância).

Após o resultado de cada palavra, conseguimos obter os substantivos mais relevantes utilizadas no Exame Nacional do Ensino Médio e o contexto em que se encontram.

Pode-se observar uma concentração das palavras em torno da temática ambiental, o que é bastante compreensível dada a importância do tema para a sociedade atual. Destaca-se, sobretudo, a palavra energia, com frequência bem superior às demais. Essa é uma pista valiosa para a compreensão das temáticas das provas, que está em harmonia com as preocupações da sociedade moderna, principalmente quando se refere à sustentabilidade ambiental. Vale lembrar que o tema "energia" é pauta governamental e área estratégica para a segurança do país. A prova reflete, mais do que uma coleção de conteúdos de nível médio, as políticas essenciais para o governo.

Um aspecto interessante é que as temáticas listadas nos substantivos remetem principalmente aos conteúdos das provas de Ciências Naturais e Ciências Humanas. Uma hipótese para explicar isso é que os conteúdos das provas de Linguagem e Matemática possuem vocabulário menos específicos e mais transversais. Levando em consideração que as provas do Enem são reconhecidamente elaboradas a partir de situações-problema, ou seja, não procuram explorar definições ou conceitos que levem à simples necessidade de memorização mas sim o raciocínio, pode-se prever que nestas provas o léxico não seja específico da área de conhecimento. Essa hipótese poderia ser testada mas foge ao escopo dessa pesquisa.

Conclusões

Os resultados demonstram a relevância temática de questões relacionadas ao meio ambiente e à sustentabilidade, pois as palavras mais frequentes remetem a esses conteúdos. Isso indica, por um lado, que a prova aborda questões essenciais para o governo. Por outro, a pesquisa acena na direção de que os estudos de nível médio poderiam voltar-se ao tratamento de temas numa perspectiva inter ou transdisciplinar, superando as limitações conteudistas.

Agradecimentos

À FAPEMIG pelo fornecimento de bolsas e auxílio financeiro.

Referências Bibliográficas

BARBOSA, M. A . **Língua e discurso: contribuições aos estudos semântico-sintáxicos**. SP: Ed.Plêiade, 1996.

BIDERMAN, M. T. **Teoria Lingüística** - (Teoria lexical e lingüística computacional), 2ªed., SP:Martins Fontes, 2001,Coleção leitura e crítica.

DAMASCENO, Elizabete Aparecida. Lexicometria e ensino de línguas. In: **Seminário do Gel**, 56., 2008, Programação... São José do Rio Preto (SP): GEL, 2008.

MILLER, M. User's Guide for VBPro: **A Program for Analyzing Verbatim Text.** University of Tennessee, Knoxville: 1995.