A produção científica do UEADSL e a linguagem R: uma análise dos principais termos citados nos anais do Congresso Nacional Universidade, EaD e *Software* Livre.

COSTA, Priscilla Tulipa[1]

RESUMO: O UEADSL é um evento *on-line,* promovido pelo LabSemiotec-UFMG, para discussão interativa das temáticas *software* livre, ensino a distância, cultura livre e suas relações com a Universidade. Este trabalho objetiva a análise comparativa, via linguagem de programação R e *software* livre RStudio, das palavras mais frequentes nos anais dos últimos três anos do evento. O resultado será a obtenção de nuvens de palavras com os termos mais citados nos trabalhos analisados e sua representatividade nesse universo de pesquisa.
Palavras-chave: *Software* livre. Linguagem R. Linguística de *corpus*. Linguística computacional.

1 INTRODUÇÃO

A experimentação ora relatada compreende uma análise comparativa da frequência das palavras nos artigos que compõem os anais do Congresso Nacional Universidade, EaD e *Software* Livre (UEADSL), evento organizado pelo LabSemiotec, da Faculdade de Letras da UFMG. Trata-se de uma exploração de dados linguísticos textuais com base na Linguística de *Corpus* e na Linguística Computacional, e foi realizado por meio da linguagem R (Versão 3.1.2) que, segundo Oushiro (2014, p. 134), é uma linguagem de programação para a análise de dados "que pode ser utilizada para realizar computações estatísticas e gráficas, compilar e anotar *corpora*, produzir listas de frequências", entre outras tarefas; do *software* livre RStudio (Versão 0.98.1087), que é um ambiente de desenvolvimento integrado para R que disponibiliza "ferramentas adicionais diretamente na interface gráfica, como a visualização dos *scripts* abertos recentemente, o histórico de linhas de comando executadas e a lista de pacotes instalados" (OUSHIRO, 2014, p. 136); e do uso de expressões regulares (ERs), que são metacaracteres usados para definir padrões de busca específicos (JARGAS, 2012).

Um dos recursos para a visualização dos resultados de medição de frequência obtidos por meio desses programas é a nuvem de palavras (*word cloud*), que consiste em um aninhamento de palavras representadas em

[1] Mestranda em Estudos Linguísticos. FALE - UFMG. priscillatulipa@gmail.com

tamanhos proporcionais à frequência com que aparecem em um conjunto de textos. Esse recurso será usado para gerar os resultados desta pesquisa.

2 OBJETIVOS

O objetivo geral é utilizar um *corpus* significativo para demonstrar de que forma o *software* livre RStudio e a linguagem R, como recursos da Linguística Computacional, podem contribuir para a análise linguística baseada em *corpus*. São objetivos específicos: 1) Formar nuvens de palavras dos artigos de cada edição dos anais do Congresso UEADSL; 2) Comparar entre si as nuvens obtidas das edições pesquisadas; 3) Observar eventuais regularidades ou substituições entre as palavras mais frequentes em cada edição.

3 METODOLOGIA

Para o estudo ora relatado foram usados quatro anais do UEADSL, de temáticas únicas, publicados no período entre 2012 e 2014 (designadamente as edições de 2012.1, 2012.2, 2013.1 e 2014.2). Após a escolha e a seleção do *corpus*, todos os textos tiveram sua formatação anulada no bloco de notas e cada artigo foi salvo separadamente, em formato txt. Ao todo, foram gerados 179 arquivos (divididos em três grupos: cada um para um ano de publicação).

A análise deu-se da seguinte forma: iniciou-se a interpretação do *script* no RStudio realizando primeiramente a limpeza da memória, a configuração do local de salvamento dos dados e do diretório de trabalho. Para a preparação do *corpus*, procedeu-se com: carregamento do pacote *tm* (*Text Mining*) no R; leitura dos arquivos do diretório; carregamento dos arquivos e criação do *corpus*; sumário dos dados; remoção dos marcadores e espaços em brancos; análise das propriedades do *corpus*; alteração das letras maiúsculas para minúsculas; remoção das *stopwords* – palavras a serem ignoradas na busca (OUSHIRO, 2014), remoção da pontuação e, por fim, dos números.

Preparado o *corpus*, procedeu-se a matriz de frequência (criação, operação e associação dos termos). Feito isso, usou-se expressões regulares para carregar os pacotes *Wordcloud*, *XML* e *RColorBrewer*, usados na formação das nuvens de palavras. Por fim, foram inseridas as linhas de comando para a criação das nuvens com os seguintes argumentos e valores: escala (8, 4),

frequência mínima (10), número máximo de palavras (Inf.), ordem randômica (False), proporção de palavras em rotação (15) e coloração das palavras (pal2). Tais comandos foram executados nos artigos coletados resultando na formação de quatro diferentes nuvens de palavras.

4 RESULTADOS E DISCUSSÃO

Da perspectiva anual individual, na figura 1 é apresentada a nuvem de palavras formada para o ano de 2012, onde verifica-se que o termo mais citado nas produções científicas daquele ano no UEADSL é “internet”, seguido por “liberdade”, “livre” e “*software*”. Para esse resultado deve-se considerar que em 2012 houve duas edições do evento, gerando, provavelmente, um número maior de artigos e, consequentemente, de palavras analisadas. Esses termos refletem os principais assuntos debatidos na temática EaD e *Software* Livre.

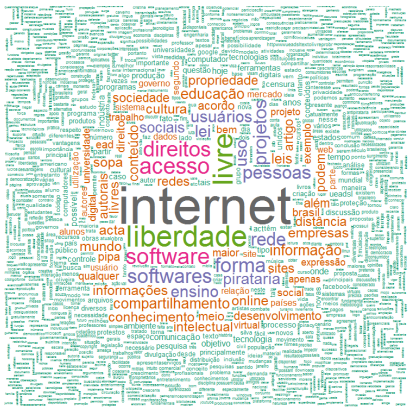

Figura 1 – Palavras mais citadas no UEADSL no ano de 2012.

Figura 2 – Palavras mais citadas no UEADSL no ano de 2013.

A nuvem criada para o ano de 2013 exibe a palavra “livre” como a mais frequente, seguida dos termos “internet”, “liberdade” e “*software*”, como mostra a figura 2. Semelhantemente ao ano anterior, as palavras “livre”, “liberdade” e “internet” são as de maior destaque. Contudo, nota-se também a frequência relevante de palavras como “educação”, “ensino” e “cultura”, que refletem assuntos debatidos nas temáticas Universidade e Cultura Livre.

A figura 3 mostra a nuvem gerada para 2014, em que as palavras mais mencionadas são “texto” (mostrada pela primeira vez como destaque principal) e “liberdade” (que se manteve frequente em relação aos anos anteriores). As outras palavras mais citadas são “análise”, “semiótica”, “forma” e “internet”, esta última com queda de frequência em relação às demais análises.

Figura 3 – Palavras mais citadas no UEADSL no ano de 2014.

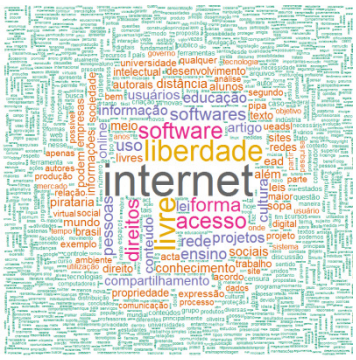

Figura 4 – Palavras mais citadas no UEADSL entre os anos de 2012 e 2014.

Por fim, em análise ampla do *corpus,* como mostra a figura 4, identifica-se que, entre 2012 e 2014, a palavra mais citada nos artigos é “internet”, seguida por “liberdade” e “livre”. Os termos “*software*” e “acesso” também se destacam na nuvem, demonstrando sua relevância para os estudos em EaD e *Software* Livre. Em relação ao destaque das palavras “livre” e “*software*”, faz-se necessário relatar que suas frequências amplas estão, provavelmente, ligadas ao termo “*software* livre”, que fora desmembrado na análise do *corpus*.

5 CONCLUSÃO

O uso da linguística computacional neste trabalho, por meio da linguagem R e do *software* livre RStudio, favoreceu o exame dos artigos publicados pelos pesquisadores de Linguística Aplicada e Tecnologia nos últimos três anos e demonstrou, imageticamente, as palavras mais citadas nos trabalhos científicos analisados, constituindo um panorama importante das informações existentes no material coletado. Da mesma forma, a pesquisa evidenciou que as palavras ressaltadas no acervo do UEADSL refletem o universo da produção acadêmica atual da área, bem como a representatividade desses termos nas diversas linhas de pesquisa sobre os temas tratados.

REFERÊNCIAS

JARGAS, Aurélio Marinho. **Expressões regulares: uma abordagem divertida**. 4 ed. São Paulo: Novatec, 2012.

OUSHIRO, Livia. Tratamento de dados com o R para análises sociolinguísticas. In: FREITAG, Raquel Meister Ko. Freitag (Org.). **Metodologia de coleta e manipulação de dados em Sociolinguística**. São Paulo: Edgard Blücher, 2014. p. 133-176.