CONTRASTANDO DUAS FERRAMENTAS PARA ANÁLISE DE CORPUS DE APRENDIZES: ANTCONC E PACOTE TM

GOMIDE, Andressa Rodrigues[1]

RESUMO: Os recursos de mineração de texto e linguística de corpus permitem o tratamento de grandes massas de texto, dando acesso a conjuntos de informações que não seriam visíveis através dos métodos tradicionais de leitura. Atualmente há um número considerável de ferramentas que permitem análise de textos. Este trabalho tem como objetivo comparar os benefícios do uso do AntConc e do pacote TextMining-R para análise da produção escrita de aprendizes de inglês de diferentes níveis de proficiência e cursos de graduação.

Palavras-chave: linguística de corpus, mineração de textos, ferramentas de concordância

1 INTRODUÇÃO

No contexto educacional, a análise de dados qualitativos é de extrema importância para a identificação de padrões que não seriam notados apenas com base em nossa intuição. Entretanto, para analisar grandes massas de textos, é necessário o domínio de ferramentas que auxiliem na mineração de textos.

O objetivo deste trabalho é analisar duas ferramentas muito utilizadas para análise de corpora de textos escritos: o software gratuito AntConc e a linguagem de programação de código-aberto R.

[1] Mestranda. FALE/UFMG gomide.andressa@gmail.com

2 FUNDAMENTAÇÃO TEÓRICA

Há um grande número de softwares de concordância e sistema de consulta on-line, tais como AntConc (ANTHONY, 2014), WordSmith Tools (SCOTT, 1996) e SketchEngine (KILGARRIFF, ADAM, et al., 2014) que apresentam uma interface considerada mais amigável do que a linguagem de programação R. No entanto, há várias boas razões para utilizar o R, dentre as quais Stefan Gries (2009) apresenta cinco. Em primeiro lugar, ele argumenta que o aprendizado e o uso de uma linguagem de programação não é tão demorado como pode parecer. Segundo ele, depois de ter desenvolvido os primeiros scripts e aprimorado algumas habilidades, é possível reutilizar os scripts, o que pode ser tão ou até mais rápido do que utilizar softwares de concordância. Além disso, o tempo de processamento requerido no R é consideravelmente menor do que o tempo exigido por estes programas. Uma segunda razão apresentada por Gries é o fato de que o usuário está no controle. Portanto, ele pode tomar decisões, tais como definir o que uma palavra é, e fazer com que seu estudo seja replicável. Não só isso, o usuário não dependerá do desenvolvedor de software. Gries apresenta como uma terceira vantagem o fato de que o R é uma linguagem de programação de código aberto, o que o torna transparente e continuamente atualizado por usuários de diversas localidades. O quarto ponto indicado por Gries está relacionado com as várias tarefas que podem ser feitas com R, em contraste com os programas de concordância. Por exemplo, o R permite a realização de avaliação estatística, anotação, recuperação de dados, representação gráfica e processamento de dados usando apenas o seu próprio ambiente. Todos estes benefícios são oferecidos gratuitamente, uma vez que o R é um software de fonte aberta, sendo esta a vantagem final apresentado por Gries.

Também oferecido gratuitamente, AntConc é outra ferramenta útil, massivamente utilizado não só por pesquisadores, mas também por estudantes e professores. De acordo com Anthony (2013), em 2012, mais de 120.000 downloads foram realizadas em mais de 80 países. Esta crescente popularidade se explica pela interface atraente amigável do AntConc, somada

a suas funções facilmente acessíveis, tais como linhas de concordância KWIC e listas de palavras-chave (ANTHONY, 2013). No entanto, AntConc não é um software de código aberto, o que faz com que a base das análises não seja totalmente transparente.

3 MÉTODO

O presente estudo utilizou como conjunto de dados uma seção do CorIsF-inglês, um corpus composto pela produção escrita dos alunos de inglês do curso presencial do programa Idiomas sem Fronteitas (IsF)[2]. O subcorpus utilizado apresenta os textos escritos pelos alunos de diferentes níveis de proficiência do núcleo do IsF na Universidade Federal de Minas Gerais (UFMG) em dois momentos: no início e ao fim da primeira metade do curso de 2014-2. As coletas foram realizadas via *GoogleForms* e os dados são salvos no formato csv, totalizando 82.858 palavras (tabela 1).

	TESTE 1		TESTE 2		TOTAL	
	Alunos	**Palavras**	**Alunos**	**Palavras**	**Alunos**	**Palavras**
MEO 2	75	6532	44	5349	119	11881
MEO 3	49	8210	33	6831	82	15041
MEO 4	77	12587	68	16149	145	28736
MEO 5	70	15603	46	11597	116	27200
TOTAL	271	42932	191	39926	462	82858

Tabela 1: distribuição do número de palavras/nível no subcorpus

Para realizar a comparação entre as duas ferramentas em questão, foram escolhidas três funções frequentemente utilizadas ao se analisar um corpus: lista das palavras mais frequentes, colocados, e linhas de concordância. Utilizando o subcorpus aqui descrito, os seguintes passos foram seguidos para que a performance das duas ferramentas fossem analisadas.

1. Limpeza e processamento dos dados

[2] http://isf.mec.gov.br/

2. Criação da lista das palavras mais frequentes
3. Identificação dos colocados mais frequentes
4. Geração das linhas de concordância

4. CONCLUSÃO

Os resultados foram agrupados na tabela abaixo (tabela 2) de forma a facilitar a leitura dos dados. Como pode-se perceber, as ferramentas apresentaram um resultado semelhante. Enquanto o uso do pacote tm na linguagem de programação R favorece o processamento e limpeza dos dados bem como as análises estatísticas, o software AntCon é ideal para o usuário que busca uma solução simples e confortável. Considerando esta interepretação, pode-se dizer que o AntCon é uma ferramenta adequada para usuários que utilizam um banco de dados já processado para análises textuais mais superficiais. Um exemplo deste público seria professores e aprendizes de línguas. Por outro lado, pesquisadores como linguístas e analistas de dados se beneficiam mais das ferramentas disponíveis na linguagem R, uma vez que esta permite uma análise mais extensa dos dados.

	Processamento			Usabilidade		Extras	
	letras maiúsculas e minúsculas	leitura de arquivos csv	padronização de palavras	link direto para KWIC	visualização agradável	lê corpus anotado	medidas estatísticas
AntConc	✓	✗	✗	✓	✓	✓	2
Linguagem R	✓	✓	✓	✗	✗	✓	várias

Tabela 2: simplificação dos resultados obtidos

REFERÊNCIAS

ANTHONY, L. A critical look at software tools in corpus linguistics. **Linguistic Research**, v. 30, n. 2, p. 141–161, 2013.

ANTHONY, L. **Developing AntConc for a new generation of corpus linguists**Corpus Linguistics Conference 2013. **Anais**...Lancaster: 2013

ANTHONY, L. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/, 2014

GRIES, S. T. H. **Quantitative Corpus Linguistics With R: A Practical Introduction**. New York: Routledge, 2009.

SCOTT, M. WordSmith Tools, Oxford: Oxford University Press. ISBN 0-19-458984-6, 1996