A análise textual exploratória como ferramenta de melhoria de um curso de formação de professores em Astronomia

Exploratory textual analysis as a tool to improve a teacher training course in Astronomy

Julio Murilo Trevas dos Santos
Universidade Federal da Fronteira Sul / Universidade Estadual de Maringá
proftrevas@gmail.com
Ana Maria Pereira
Fundação Parque Tecnológico Itaipu
ana.pereira@pti.org.br
Janer Vilaça
Fundação Parque Tecnológico Itaipu
janer@pti.org.br
Neide Maria Michelan Kiouranis
Departamento de Química, Universidade Estadual de Maringá
nmmkiouranis@gmail.com

Resumo

Este trabalho propôs uma metodologia de análise textual exploratória aplicada às respostas de avaliação de um curso de capacitação de professores na área de Astronomia. O objetivo foi determinar indicadores para o aprimoramento do próprio curso. A análise textual é exploratória por situar-se na interseção da mineração de dados, da linguística de corpus e da análise de conteúdo. A metodologia foi apoiada pelo uso do *software* de código aberto *Voyant Tools*. A partir da codificação e categorização dos enunciados das questões e processamento no *Voyant*, conseguiu-se categorizar as respostas e identificar tendências que permitem determinar alguns indicadores para o aprimoramento do curso.

Palavras chave: análise textual exploratória, análise de conteúdo, mineração de texto, avaliação, educação em ciências, formação de professores

Abstract

This work proposes an exploratory textual analysis methodology applied to the evaluation responses of a teacher training course in the area of Astronomy. The objective was to determine indicators for the improvement of the course itself. Textual analysis is exploratory because it lies at the intersection of text mining, linguistics of corpus and content analysis. The methodology was supported by the use of open-source software Voyant Tools. From the coding and categorization of the questions and Voyant's processing, we were able to

categorize the answers as well as identify trends that allow us to determine some indicators for the improvement of the course

Key words: exploratory textual analysis, content analisys, text mining, evaluation, science education, teacher training

Introdução

Na Fundação Parque Tecnológico de Itaipu (PTI) está instalado o Polo Astronômico Casimiro Montenegro Filho (Polo), um centro de ciências que desenvolve atividades educacionais e de pesquisa focadas na Astronomia. Dentre as atividades educacionais cita-se um curso de fundamentos de ensino de Astronomia para professores do Ensino Fundamental (Curso). É um curso que atendeu a 1.681 professores nos últimos 8 anos. Primando pela qualidade do Curso, a equipe responsável elabora e aplica diferentes instrumentos de avaliação do/no curso. O Curso vem atendendo um número expressivo de professores, assim tem sido gerado uma grande quantidade de dados nas avaliações. Tal quantidade exige uma abordagem de processamento e análise automatizada para agilizar as tomadas de decisão em relação ao Curso.

O desafio, portanto, era estabelecer uma metodologia de análise assistida por computador capaz de gerar indicadores de aprimoramento do Curso. Como as avaliações possuem questões dissertativas, uma metodologia de análise de texto se configurou a mais adequada.

Um princípio foi a investigação no texto de dados e informações desconhecidas. Isso se aproxima do conceito de mineração de textos, um "[...] processo de extração de informações de interesse e padrões não-triviais ou descoberta de conhecimento em documentos de texto" (ARANHA, PASSOS, 2006, p.2) não estruturados ou semiestruturados. Simultaneamente há aproximação às técnicas de Linguística de *Corpus* (LC), nas quais ferramentas computacionais são utilizadas para extrair e organizar informações do *corpus*[1] (MELLO; SOUZA, 2012, p.5). Por outro lado, também há aproximação à Análise de Conteúdo (AC), compreensível como uma metodologia para produzir inferências de um texto, identificar sistematicamente as mensagens implícitas no texto (BAUER, 2015, p.192). Bauer (2015, p.194) cita diferentes tipos de pesquisa de AC, desde o mais simples, a determinação de frequência dos dados codificados, ao mais interessante, análise normativa que efetua comparações com padrões. Percebe-se que uma metodologia de AC acaba envolvendo os mesmos recursos computacionais da mineração de texto e da Linguística de *Corpus*.

Buscou-se, portanto, uma análise textual inspirada nas técnicas de AC, LC e mineração, que, portanto, foi denominada de exploratória por não atender as características de análises mais complexas.

Metodologia

Trata-se de uma pesquisa qualitativa, com procedimentos qualitativos e quantitativos, que envolve a análise textual exploratória de *corpora* – conjuntos de respostas às avaliações do Curso. Foram analisadas as avaliações do Curso realizado em setembro de 2018, do qual participaram 43 professores de uma rede pública municipal de educação do oeste paranaense.

[1]Entendendo-se *corpus* como um conjunto textual "[...] constituído de dados autênticos, legíveis por computador e representativos de uma língua ou da variedade da língua a qual se deseja estudar. " (MELLO; SOUZA, 2012)

O Polo aplicou três avaliações ao longo do curso. A primeira avaliação ocorreu 30 dias antes do início do curso, a segunda no meio do curso e a terceira ao final do curso. As avaliações foram elaboradas e disponibilizadas como questionários em formato digital através do serviço *web* Formulários *Google* (GOOGLE, 2018). O preenchimento dos formulários foi um dos requisitos para a certificação no curso.

Os questionários continham questões fechadas, semifechadas e abertas (semifechadas as questões com respostas pré-definidas, mas com solicitação de justificativa dissertativa). O Polo elaborou um relatório para cada avaliação. Nos relatórios as respostas de todos os professores foram agrupadas por questão. Para as questões abertas foram listadas sequencialmente as respostas de cada respondente. As respostas agrupadas por questão aberta nesses relatórios foram eleitas os *corpora*[2] deste trabalho.

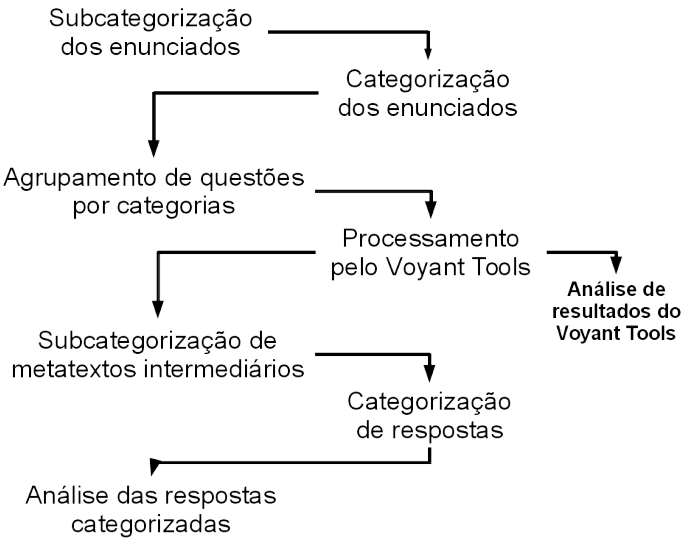

Figura 1: Esquema simplificado da análise textual exploratória das avaliações do curso de Astronomia

O processo de análise foi composto por sete etapas principais (Figura 1). Na primeira, cada avaliação foi rotulada: primeira avaliação, A1; segunda avaliação, A2; e última A3. Os enunciados das questões foram destacados e organizados por avaliação. Em uma leitura livre[3] evidenciou-se o tema mais significativo, depois denominado subcategoria, para cada enunciado. Para cada subcategoria inferiu-se a característica ou temática mais ampla a qual está relacionada a subcategoria.

Na segunda etapa as subcategorias foram agrupadas segundo natureza comum - categoria de enunciado. Organizou-se os dados da categorização dos enunciados, relacionando-se a categoria, suas subcategorias, as questões associadas a cada subcategoria e os indicadores respectivos. Os indicadores são excertos dos enunciados que evidenciam as subcategorias associadas.

Desse modo as questões foram agrupadas por subcategoria. Conseguiu-se dessa maneira formar conjuntos de questões, de diferentes avaliações, que deveriam estar relacionadas. Isto significou o estabelecimento de um critério para identificar e justificar quais respostas (de quais questões) de uma avaliação poderiam ser cotejadas com as de outra. Portanto, a análise das respostas foi efetuada por grupo de questões segundo a subcategoria a qual pertenciam.

[2] *corpora* é o plural e *corpus*.

[3] Neste trabalho, leitura livre é uma leitura em primeiro contato, interpretativa, desprovida de pré-indicadores e regras.

Neste trabalho selecionou-se uma subcategoria para processamento e análise, priorizando-se aquela que transportava as percepções de potencialidades, fragilidades e impactos do Curso. Após a seleção da subcategoria de enunciados, o *corpus* de cada questão foi inserido no *software Voyant Tools* (SINCLAIR; ROCKWELL, 2016). O *software* é um projeto de código aberto, gratuito, independente de plataforma e com interface amigável, o que amplia as possibilidades de pesquisa na área de educação em ciências. O *Voyant* pode ser classificado como um concordanciador e frequenciador, um *software* capaz de contar palavras, buscar palavras ou expressões específicas, selecionar os contextos de ocorrência das palavras, organizar sequências de palavras segundo a frequência (MELLO; SOUZA, 2012). Ele também é capaz de calcular correlações e vínculos entre palavras, efetuar diferentes representações gráficas, entre outros recursos.

O *software* não é configurado para reconhecer a sintaxe e gramática da língua portuguesa. Faz-se necessário configurá-lo para remover aquelas palavras que são irrelevantes na análise, como artigos, pronomes, entre outras. Essas palavras são denominadas *stop words*, ou palavras vazias. Utilizou-se uma lista pré-definida de *stop words* da língua portuguesa disponibilizada no GitHub[4]. Em algumas situações uma palavra considerada vazia pode ser relevante em um *corpus* específico, por isso gerou-se um conjunto de palavras de maior frequência com *stop words* e outro sem *stop words*. Contudo os resultados dos outros recursos foram obtidos em função do conjunto sem as *stop words*.

Para cada *corpus* registraram-se os seguintes dados processados pelo *Voyant Tools*: a) número total de palavras e número de palavras em forma única (sem repetições); b) a densidade de vocabulário; c) a média de palavras por sentença; d) as palavras e respectivas frequências (com e sem *stop words*); e) algumas expressões e suas frequências; f) a nuvem de palavras; e g) o diagrama de vínculos. A densidade de vocabulário é a razão do número de palavras em forma única pelo número total. A densidade pode ser um indicativo da complexidade do texto, pois quanto maior o seu valor menor é o número de palavras repetidas. Selecionou-se para cada *corpus* em torno de 15 palavras com maior frequência. No caso das sequências de palavras, o critério de seleção foi: a expressão precisava ser relevante, ter significado (por exemplo, "um tempo maior" foi considerada relevante, enquanto "muito bem" sem relevância). O diagrama de vínculos mostra conexões entre palavras

Cada palavra, dentre as mais frequentes sem *stop words*, foi avaliada em seu contexto, que é o trecho de texto onde está presente, considerando também as expressões frequentes. Dessa avaliação extraíram-se possíveis ideias transmitidas. Um segundo conjunto de ideias foi obtido da avaliação do diagrama de vínculos. Um terceiro conjunto foi obtido da leitura livre das respostas. Nessa leitura buscaram-se ideias que não emergiram da frequência de palavras e expressões. Todos os resultados da análise com o *Voyant* foram tabulados para a execução da etapa subsequente.

A quinta etapa foi inciada pela leitura crítica dos três conjuntos de ideias, associado a cada *corpus*, e elaboração de sínteses, os metatextos intermediários. Uma vez tabulados, efetuou-se a leitura dos metatextos para identificação de temas emergentes, as subcategorias de respostas (Figura 2).

[4]A lista de *stop words* disponibilizada em arquivo em formato texto (".txt") no endereço <https://gist.github.com/alopes/5358189>.

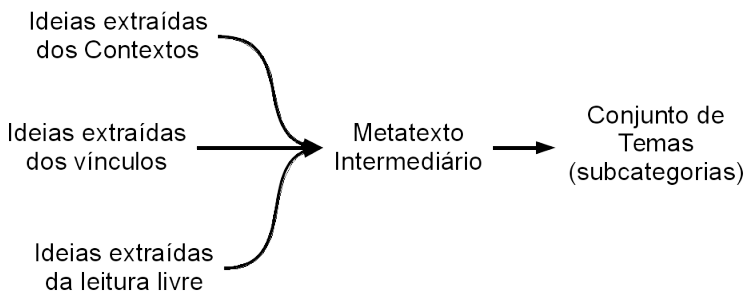

Figura 2: esquema de subcategorização de respostas

Algumas subcategorias se repetem entre os *corpora*. Analisaram-se as subcategorias na busca de aproximações entre elas que permitissem o agrupamento em temas mais amplos, as categorias de respostas. Deve-se destacar que a rotulação de subcategorias e categorias, em todas as etapas, é um processo iterativo e de refinamento, cujos resultados ainda são passíveis de ajustes. Esta penúltima etapa, de categorização de resposta culminou com a elaboração de uma tabela relacionando *corpus*, categorias e número de subcategorias associadas. A última etapa configurou-se em uma análise interpretativa dos *corpora* a partir de suas categorizações.

Resultados e Discussões

A avaliação A1 foi composta por oito questões, A2 por dez questões e A3 por dezoito. O elevado número em A3 foi justificado pela inserção de questões específicas a uma aula do Curso. A avaliação A1 coletou dados das expectativas dos professores, as avaliações A2 e A3 coletaram dados das percepções sobre o curso. Cada questão foi rotulada com seu número de ordem e com o rótulo da avaliação (exemplo a quarta questão da primeira avaliação foi rotulada "A1.4"). Após uma leitura livre permitiu a inferência das subcategorias: autoavaliação; formação docente em Astronomia; participação no curso; perfil docente; aspectos conceituais; fontes de pesquisa; concepções alternativas; características gerais; conteúdos de Astronomia; corpo docente; aspectos procedimentais; organização de conteúdos; e recursos didáticos. As subcategorias autoavaliação, formação docente em Astronomia, participação no curso e perfil docente referem-se a como cada indivíduo se autoavalia, a seu perfil e formação, por isso foram agrupadas na categoria Aspectos Individuais. As subcategorias aspectos conceituais, fontes de pesquisa e concepções alternativas referem-se à utilização de fontes de pesquisa, à percepção de concepções alternativas de alunos e às mudanças conceituais, sendo agrupadas na categoria Aspectos Educacionais. As subcategorias restantes referem-se ao Polo Astronômico, ao corpo docente, às potencialidades e fragilidades, entre outras características do Curso, o que sugeriu o agrupamento na categoria Aspectos Institucionais.

Selecionou-se então a subcategoria "características gerais", por essa transportar informações sobre as potencialidades e fragilidades do curso. Logo a relação entre categoria, subcategoria, questões e indicadores é apresentada no Quadro 1. As questões da subcategoria selecionada envolvem o registro de aspectos positivos, negativos, potencializadores, impeditivos e sugestões, o que ratifica a coerência do agrupamento. Concomitantemente, compreende-se que por serem aspectos de ações passadas, não há no grupo qualquer questão de A1 (diagnóstico de ações futuras). Porém há a possibilidade de correlação com uma questão de A1.

CATEGORIA DE ENUNCIADO	SUBCATEGORIA DE ENUNCIADO	QUESTÕES	INDICADORES
Aspectos Institucionais	características gerais	A2.8, A2.9, A2.10, A3.11, A3.12, A3.14, A3.15	"registre os aspectos positivos"; "registre aspectos negativos"; "indique sugestões ou ideias"; "registre os aspectos potencializadores e motivadores"; "registre os aspectos impeditivos"; "Você classifica o curso..."; "para o curso ser inovador [...] o que você sugere"

Quadro 1: Questões e indicadores associados à subcategoria características gerais.

Os *corpora* das questões mencionadas foram processados individualmente no *Voyant*. Exemplos de resultados são apresentados nas Figuras 3 e 4, bem como nos Quadros 2 e 3.

Figura 3: Nuvem de palavras obtida no processamento do corpus de A3.11

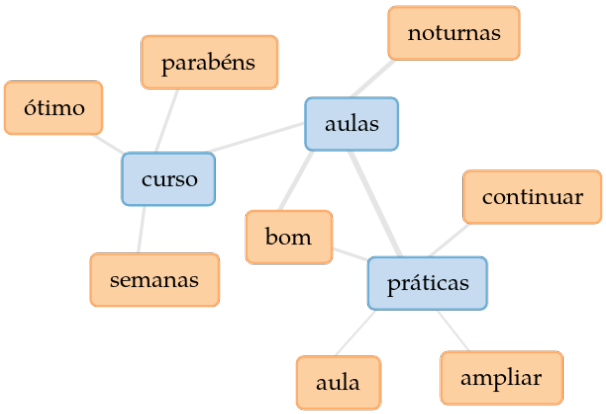

Figura 4: Diagrama de vínculos obtido no processamento do corpus de A2.10

Em uma leitura livre e flutuante (não restrita a sentidos pré-definidos) do diagrama, pode-se inferir a continuidade e ampliação de aulas práticas. No mesmo diagrama confirma-se a limitação do *software* em relação à sintaxe e gramática ao considerar "aula" e "aulas" como palavras diferentes.

Corpus	Sumário de contagem de palavras	Palavras mais frequentes COM Stop Words	Palavras mais frequentes SEM Stop Words	Expressões Frequentes
A2.9	Total de 415 palavras e 198 formas únicas Densidade de vocabulário de 0.477 Média de palavras por sentença de 13.0	não (20) para (14) tempo (12) a (10) que (10) de (9) o (9) negativos (8) as (7) curso (7) pouco (7) aspectos (6) do (6) em (6) um (6)	tempo (12) negativos (8) curso (7) pouco (7) aspectos (6) aulas (5) conteúdo (5) curto (4) negativo (4) nenhum (4) alguns (3) dias (3) informações (3) muita (3) nada (3)	"não tenho aspectos negativos" (3) "muitas informações em" (2) "pouco tempo para" (2);

Quadro 2: Exemplo de resultados do processamento e análise com Voyant Tools

Corpus	ideias extraídas dos contextos (palavras frequentes)	ideias extraídas do diagrama de vínculos (palavras frequentes)	ideias extraídas de leitura livre das respostas
A2.9	pouco tempo para muitas informações e para estudar não há aspectos negativos o curso necessita de carga horária maior aulas em tempo integral cansativas;	tempo curto	conteúdos difíceis que requerem aprofundamento artigos disponibilizados considerados "muito técnicos";

Quadro 3: Continuação do exemplo de resultados do processamento e análise com Voyant Tools

Analisando o sumário de contagem de palavras verifica-se que entre a segunda e terceira avaliações houve um aumento no número de palavras (Figura 5), um pequeno aumento na média de palavras por sentença, mas sem alteração significativa na densidade de vocabulário. Como não houve aumento na complexidade dos textos, sugere-se um receio ao responder a segunda avaliação. A densidade de vocabulário é um resultado que deve ser ressaltado pelo potencial de indicar alterações nas capacidades de reflexão e de pensamento crítico.

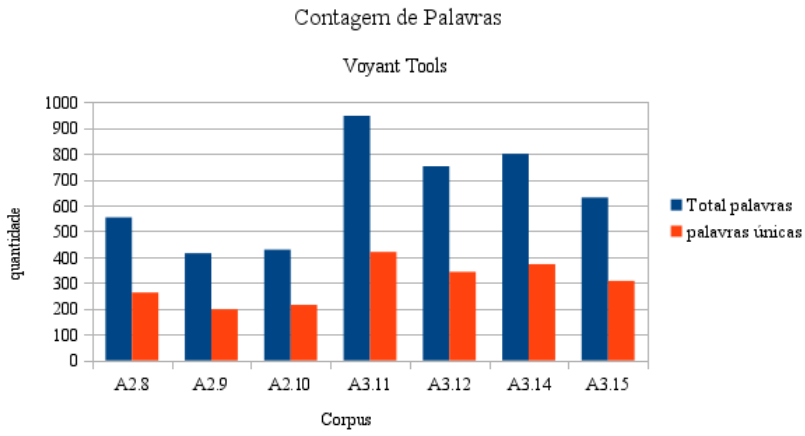

Figura 5: Contagem de palavras dos *corpora* processados no *Voyant Tools*

O Quadro 3 exemplifica conjuntos de ideias para um *corpus*. Esses conjuntos propiciaram a elaboração de metatexto intermediário e de subcategorias de respostas (Quadro 4).

Corpus	**Metatexto Intermediário**	**Subcategorias**	**Número de subcategorias**
A2.9	o curso apresentou conteúdos difíceis, os quais requerem aprofundamento; a quantidade de informações foi elevada para o (curto) tempo do curso, indicando a necessidade de aumento de carga horária; o desenvolvimento dos encontros em período integral são fatigantes; os artigos (científicos) disponibilizados foram considerados difíceis ("muito técnicos")	aprofundamento conceitual carga horária texto técnico-científico jornada de curso	4

Quadro 4: Exemplo de metatextos intermediários e de subcategorias de respostas

Da elaboração de metatextos para todos os *corpora*, propuseram-se as seguintes subcategorias de respostas: aprofundamento conceitual; atividades práticas; aulas dinâmicas; carga horária; coerência conceitual; complexidade conceitual; contextualização; domínio de conteúdo; formação continuada; formação de conceitos; natureza da ciência; nível educacional; observação astronômica; estrutura de curso; jornada de curso; pesquisa; processo de ensino e aprendizagem; recursos didáticos; texto técnico-científico.

A reorganização das subcategorias sob temáticas mais abrangentes produziu as categorias de respostas:

a) Domínio Conceitual – envolve aspectos relativos aos conceitos de Astronomia
b) Domínio Estrutural – envolve aspectos relativos à estrutura e à organização do curso
c) Domínio Científico – envolve noções de pesquisa, ideias e concepções sobre ciência e sobre contextualização
d) Domínio Educacional – envolve aspectos de formação docente, de processos de ensino e de aprendizagem
Em outra representação gráfica (Figura 6), dispôs-se a representatividade das categorias de respostas nos *corpora* em função do número de subcategorias.

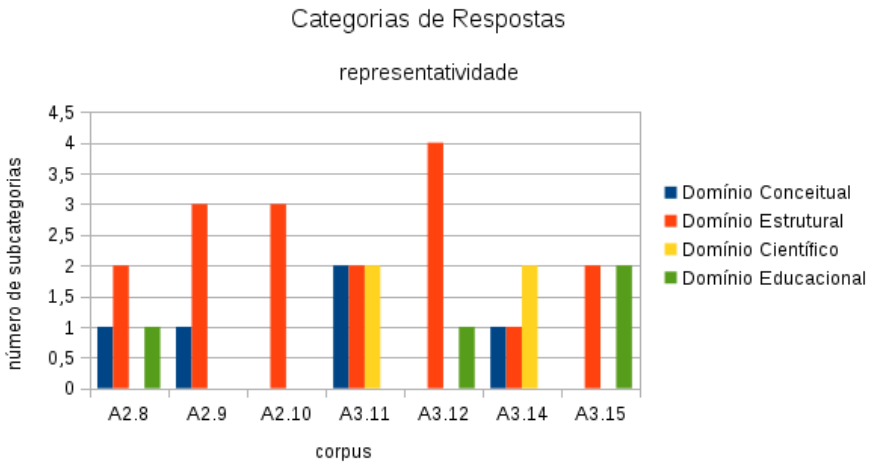

Figura 6: Representatividade das Categorias de Respostas nos *corpora*

A Figura 6 indica uma tendência, as repostas dos professores abordaram predominantemente os aspectos de estrutura e organização do Curso. Minoritariamente, foram abordadas ideias sobre pesquisa, natureza da ciência e contextualização. E no intermédio encontram-se os

aspectos conceituais, de formação docente e de ensino aprendizagem. Esse resultado é parcial, pois tratam-se de categorias de respostas vinculadas a uma subcategoria de enunciados. Somente o confronto com as categorias de respostas das demais subcategorias de enunciados poderá ampliar as interpretações. Um segundo fator é a possibilidade de explorar outras combinações dos mesmos dados/resultados. Um terceiro fator é a possibilidade de análise de outros instrumentos diagnósticos aplicados durante o curso.

Considerações Finais

O resultado, embora parcial, fornece bons indicadores para a reorganização do Curso, de modo que sejam promovidas relações da educação em Astronomia com a epistemologia do professor, a epistemologia das ciências, a história das ciências, o desenvolvimento de capacidades de pensamento crítico, entre outros aspectos. E uma das ações resultantes deste trabalho será a proposição de implementações em nível de assuntos, atividades e avaliações no Curso. Considerando o número de atendimentos no Curso, do ponto de vista metodológico, o processo de análise textual exploratória pode ser considerado efetivo no tratamento de grande quantidade de dados. Isto é importante para tomadas de decisão mais ágeis e melhor direcionadas aos alvos prioritários.

Os resultados satisfatórios induzem à ampliação dos *corpora*, quer pela análise de outras subcategorias de enunciados, quer pela análise de outras turmas participantes, ou pela participação do mesmo grupo de professores em atividades de formação posteriores organizadas pelo Centro.

Referências

ARANHA, Cristian; PASSOS, Emmanuel. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**. v. 5, n. 2, 2006, p.1-8. Disponível em <http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171>. Acesso em 12 out 2018. DOI: https://doi.org/10.21529/RESI.2006.0502001

BAUER, Martin W. Análise de Conteúdo Clássica: uma revisão. In: BAUER, Martin W.; GASKELL, George (orgs.). **Pesquisa qualitativa com texto, imagem e som**: um manual prático. 13 ed. Petrópolis: Vozes, 2015.

GOOGLE. **Formulários Google**. Disponível em <https://www.google.com/forms/about/>. Acesso em 14 out 2018.

MELLO, Heliana; SOUZA, Renato. A Linguagem da Ciência: prospecção de dados baseados em corpora. **STIS Seminários Teóricos Interdisciplinares do SEMIOTEC - Cadernos Didáticos e Anais**. v.1, n.1, 2012, 19p. Disponível em <http://www.periodicos.letras.ufmg.br/index.php/stis/issue/view/177>. Acesso em 11 out 2018.

SINCLAIR, Stéfan; ROCKWELL, Geoffrey. **Voyant Tools**. Web. 2016. Disponível em <https://voyant-tools.org/>. Acesso em 07 out 2018.