MINERAÇÃO DE TEXTO NO APOIO DA ESCRITA ACADÊMICA

TEXT MINING IN SUPPORT OF ACADEMIC WRITIING

Dr. Eliseo Reategui, Universidade Federal do Rio Grande do Sul -UFRGS, eliseoreategui@gmail.com
Me. Simone de Oliveira de Emer, Universidade Federal do Rio Grande do Sul -UFRGS, simone.emer@gmail.com
Esp. Jocimara de Lima Mauer, Faculdade da Serra Gaúcha – FSG, jocimara_mauer@hotmail.com
Esp. André Gomes, Faculdade da Serra Gaúcha – FSG, andre.gomes@fsg.br
Tecnólogo Leandro Leonel Dapper, Faculdade da Serra Gaúcha – FSG, leandro.dapper@fsg.br

Resumo: Esta pesquisa trata de investigar como uma ferramenta de mineração de texto, capaz de identificar incoerências, poderá apoiar no ensino e na aprendizagem, no processo de construção da escrita acadêmica científica, para qualificar a produção textual dos trabalhos de conclusão de curso na Faculdade da Serra Gaúcha. Para isso, utilizou-se como fundamentação teórica as teorias de produção textual e mineração do texto. A aplicação dessa pesquisa está ocorrendo como uma investigação contínua semestral, pois faz parte de uma tese de doutorado. Este artigo foi elaborado para exemplificar o que e como se pretende contribuir para qualificar a produção textual, no entanto, ele apresenta apenas o primeiro semestre de pesquisa, assim como, a organização inicial dessa proposta.

Palavras-Chave: Produção Textual, Trabalho de Conclusão de Curso, Mineração de texto, Sobek.

Abstract: This research is to investigate how a tool for text mining capable of identifying inconsistencies can support teaching and learning in the construction of scholarly scientific writing process to qualify the textual production of course conclusion papers at Faculdade da Serra Gaúcha. For this, we used as a theoretical foundation, theories of text production and text mining. The application of this research is occurring as a semiannual continuous research, as part of a doctoral thesis. This article was written to illustrate what and how we intend to contribute to qualify textual production, however, it presents only the first semester of research as well as, the initial organization of this proposal.

Keywords: Textual Production, Course Conclusion Paper, Text Mining, Sobek.

1. INTRODUÇÃO

A escrita é uma das formas mais antigas de representação da comunicação humana, desde os homens primitivos que deixaram o relato do seu cotidiano em diversos registros históricos até os tempos contemporâneos que se vive do registro digital, da era do *twitter*, *facebook*, *hiper*textos colaborativos, *chats*, entre outros conceitos que hoje fazem parte da rotina social, acadêmica ou profissional.

Apesar do contexto interativo e inovador vivenciado pelos acadêmicos, é possível perceber as dificuldades dos alunos para escrever, mesmo tendo em consideração um grande contato com a escrita cotidiana por meio destes recursos comunicacionais citados, uma parcela considerável dos acadêmicos, quando ingressam no ensino superior se deparam com a dificuldade de escrever, de como transformar os

seus pensamentos em palavras formais, na estrutura de um texto aceito cientificamente para o ensino superior.

No ensino superior a produção textual é uma prática rotineira comum e exigida por todos os cursos da graduação ou pós-graduação, o que irá variar, é a escala de exigência de quantidade ou qualidade dessa escrita. Mas, ela irá acompanhar o acadêmico do ingresso até o término do seu curso. Assim, no universo acadêmico, o aluno irá produzir uma grande diversidade de textos, desde resumos simples, resenhas, fichamentos até artigos científicos para publicação, relatórios técnicos e monografias, no nível de complexidade que cada etapa, disciplina ou estágios exigirá.

O texto é apresentado como uma construção de signos, que podem ser elaborados com auxílio das estruturas internas ou externas por um sujeito, para expressar o seu pensamento, suas opiniões, suas concepções, seus valores ou crenças, que surgem do processo de interação com outros sujeitos a partir da experiência num determinado tempo e contexto histórico. E se esse sujeito revela as suas interações por meio da escrita, é porque deseja se comunicar, ou seja, alguém para ler o comunicado (VYGOTSKY, 1987; BAKHTIN, 2000).

E neste processo, muitos acadêmicos apresentam e relatam as suas dificuldades, pois sabem o que querem expressar, sabem para quem desejam comunicar as suas ideias, mas o maior empecilho nesta dinâmica é como fazer isso de forma clara, coerente, concisa e conecta, conforme o que se está sendo exigido no ensino superior.

O objetivo geral da pesquisa é investigar como uma ferramenta de mineração de texto, capaz de identificar incoerências, poderá apoiar no ensino e na aprendizagem, no processo de construção da escrita acadêmica científica, para qualificar a produção textual dos trabalhos de conclusão de curso.

Para organizar, a investigação e a coleta se estrutura a partir dos seguintes objetivos específicos: a) Explorar a ferramenta de mineração de texto SOBEK; b) Construir a metodologia de aplicação da ferramenta; c) Elaborar e validar os critérios de análise textual com base teórica e aplicação prática; d) Aplicar o experimento com a ferramenta para formandos de diversos Cursos da Faculdade da Serra Gaúcha. Este projeto de pesquisa também será enviado para análise e submissão do comitê de ética.

Esta pesquisa organiza-se para apoio ao desenvolvimento da Escrita Acadêmica Científica, conforme as características que esse processo exige em sua produção, levando em consideração a coerência, a coesão e conectividade, nesta produção teórica os conceitos propostos serão abordados de modo sucinto devido a fase inicial e experimental da pesquisa. Diante disso, levanta-se a seguinte hipótese:

a) O acadêmico e o professor orientador repensarão o seu texto a partir da aplicação do método de análise textual com o uso da ferramenta de mineração de texto?

Espera-se que a unidade de aplicação do experimento entre o método de uso da ferramenta, a mediação pedagógica e o uso de critérios de avaliação textual resultará na qualificação do TCC dos acadêmicos envolvidos no processo.

A partir da situação problema apresentada, referente às dificuldades na produção da escrita acadêmica relatada por uma parcela significativa de alunos da Faculdade da Serra Gaúcha (FSG) na cidade de Caxias do Sul (RS), justifica-se o presente estudo de como apoiar o acadêmico no processo da elaboração de textos científicos, principalmente quando se trata de uma fase muito especial em sua trajetória acadêmica, na elaboração do seu Trabalho de Conclusão de Curso (TCC).

2. REFERENCIAL TEÓRICO

Este referencial teórico apresentado propõe o uso de teorias que compreendem a produção textual acadêmica, as tecnologias educacionais e a mineração de texto para fundamentar a pesquisa pretendida.

A importância da escrita como uma evolução da humanidade é algo indiscutível, conforme Lévy (1998) as sociedades que se desenvolveram pela oralidade, precisaram da memorização, mas a partir da escrita exigiram-se formas mais complexas de pensar e registrar o pensamento.

Corrobora Bakhtin (2000) quando afirma que a atividade mental se exprime exteriormente por meio de signos, como também ela, só existe sob a forma de signos. Vygotsky (1987) relaciona o desenvolvimento do pensamento e da linguagem por meio dos signos e instrumentos, como forma de mediação entre o sujeito e o objeto de aquisição da aprendizagem. Estes aspectos necessitam ser levados em consideração para elaboração de um texto acadêmico, de que forma e quais as características que irão compô-lo nesta dinâmica.

Para análise de critérios acadêmicos de um texto formal, sugerem Beaugrande e Dresler (1983) que a coerência e a coesão são os principais pontos. A coerência define-se como os conceitos (conteúdos) do texto, o que dará sentido cognitivo. A coesão são os mecanismos gramaticais e lexicais que compõem o texto. A unidade e o equilíbrio das duas promove a interrelação semântica entre os elementos do discurso, o que pode ser chamado de conectividade textual.

Apesar da realidade presente, dos avanços tecnológicos, muitos acadêmicos encontram-se com uma significativa defasagem na leitura e na escrita, geralmente advindos de uma educação básica deficiente (RIOLFI; IGREJA, 2010). Hoje as Instituições de Ensino Superior (IES) buscam alternativas para minimizar estas situações por meio de disciplinas específicas, como Português, Metodologia da Pesquisa Científica, ou com oficinas, que algumas intitulam nivelamento ou reforço acadêmico. Contudo, é preciso salientar o papel do professor nesse universo de orientação para produção textual científica e as tecnologias educacionais, as quais poderão apoiar significativamente neste processo.

Nesse sentido, é relevante destacar as transformações tecnológicas que se apresentam na realidade atual e estas conforme Lévy (1998) são frutos de uma sociedade e de uma cultura. Novas maneiras de pensar e de conviver estão sendo elaboradas no mundo das telecomunicações e da informática. As relações entre os homens, o trabalho, a própria inteligência dependem, na verdade, da metamorfose incessante de dispositivos informacionais de todos os tipos. Escrita, leitura, visão, audição, criação, aprendizagem são capturados por uma informática cada vez mais avançada. Para ele o texto sempre foi virtual, ele afirma que quando ocorre a leitura, o sujeito se atualiza do assunto e constrói uma paisagem semântica e móvel.

Contribui Ausebel (2003) quando afirma que o conhecimento prévio do aluno é a chave para aprendizagem significativa. Por isso, se torna relevante utilizar o próprio conhecimento do acadêmico referente ao seu contexto e interesse tecnológico para aprender. Conforme o autor, quanto mais se sabe, mais se aprende. Seus conceitos são compatíveis com o desenvolvimento cognitivo de Piaget (1986) e com o sociointeracionismo de Vygotsky.

Contudo, corrobora Lévy (1998) que essa nova educação deve refletir a imagem livre de espaços de conhecimentos emergentes, abertos, contínuos, em fluxo, não-lineares, se reorganizam de acordo com os objetivos ou os contextos, nos quais cada um ocupa uma posição singular e evolutiva. Assim, Kenski (2007) relata que a maneira de

ensinar e aprender mudou muito desde que as Tecnologias de Comunicação e Informação (TICs) começaram a se expandir na sociedade. Já Warschauer, Meskill (2000) complementa a mediação do professor e seu potencial crescimento, quando explana que projetos educacionais que envolvem tecnologia, podem oferecer condições para o professor aprender, refletir, a sua atuação pedagógica com novas ferramentas.

A partir do estudo apresentado das necessidades de investigar e apoiar os acadêmicos na construção da escrita científica, das transformações tecnológicas atuais e do papel do professor como mediador nesta dinâmica, é possível, apresentar como um recurso tecnológico viável o método de mineração de texto para contribuir no processo de produção da escrita no ensino superior.

Nesse sentido, a mineração de texto pode ser definida como um processo intensivo de conhecimento no qual um usuário interage com uma grande quantidade de documentos utilizando ferramentas para análise dos mesmos. Os sistemas de mineração baseiam-se em rotinas de pré-processamento, algoritmos para descoberta de padrões e elementos para apresentação dos resultados (FELDMAM, SANGER, 2007).

Uma técnica bastante comum utilizada é a representação das características de um documento por meio de um modelo de espaço vetorial. Nesta técnica, cada termo do documento torna-se uma característica dimensional. O valor de cada dimensão pode indicar o número de vezes que o termo aparece no texto, ou pode indicar o peso do termo a ser considerado (SCHENKER, 2003).

Os grafos gerados a partir da aplicação de uma ferramenta de mineração de texto são, conforme Schenker (2003) construções matemáticas importantes e efetivas para realizar a modelagem de relacionamentos e de informação estrutural. Esta técnica de mineração de textos utilizando grafos descobre as palavras com maior ocorrência no texto e identifica se elas estão próximas (REATEGUI, 2011).

3. METODOLOGIA DA PESQUISA

A pesquisa está ocorrendo a cada semestre, vários experimentos vêm sendo realizados para verificar a validade da utilização, pois o grupo de pesquisa envolvido na investigação pretende de fato utilizar a ferramenta de mineração de texto Sobek de modo institucional. O qual é caracterizado por uma ferramenta de mineração de texto com o propósito de investigar as características de uma escrita cientifica para qualificar a produção textual acadêmica de Trabalho de Conclusão de Curso (TCCs) no sentido de análise da coesão, coerência e conectividade. Devido à fase inicial da pesquisa a ferramenta ainda está em fase experimental.

Nesse artigo será relatado o experimento realizado durante um semestre na Faculdade da Serra Gaúcha com 20 trabalhos que foram submetidos à análise.

A investigação ocorreu a partir de uma pesquisa exploratória, com abordagem qualitativa, com o procedimento de pesquisa documental e observação e aplicação da análise de conteúdo.

A metodologia se deu em duas etapas:

1ª Etapa: Elaboração e validação dos critérios de análise textual (teórico e prático) para o uso da ferramenta. Estes critérios foram elaborados com base nas pesquisas de Beuagrande e Dresler, Motta, Garcia e Bakhtin. Numa perspectiva de critérios essenciais para elaboração de um texto acadêmico.

Seguem esses critérios como:

a) Coerência: Conceitos (conteúdos) do texto, o que dará sentido cognitivo.

Continuidade, progressão, não contradição interna ou externa, articulação com presença ou pertinência. O qual apresenta a ligação com as ideias ou procedimentos, o que dará sentido a produção textual.
b) Coesão: Conforme texto da Infoescola são mecanismos linguísticos que permitem uma sequência lógico-semântica entre as partes de um texto, sejam elas palavras, frases, parágrafos ou outros. No entanto, nesta pesquisa, o Sobek utiliza-se de dados quantitativos, por isso, inicialmente se delimitará no uso dos mecanismos gramaticais e lexicais que compõem o texto; e
c) Conectividade: A unidade e o equilíbrio das duas promove a interrelação semântica entre os elementos do discurso, o que pode ser chamado de conectividade textual.

2ª Etapa: Aplicação do experimento

A técnica de coleta aplicada para abordagem qualitativa foi a técnica de relatórios e diário do pesquisador gerados a partir da submissão dos textos (TCCs) no laboratório de informática nas aulas de orientação da monografia. Sendo que, estes foram analisados por meio da técnica de análise de conteúdo. Conforme Bardin (2002) organiza-se pela ausência ou presença de determinadas características de conteúdo na coleta, por meio de categorias, subcategorias e inferências.

Diante disso, a pesquisa pretendida possui um campo empírico de atuação a partir de situações reais e necessárias de intervenção para qualificar a produção textual de futuros profissionais no mercado de trabalho, assim como, futuros pós-graduandos em outros níveis de ensino. Esta contribuição cognitiva planejada para os acadêmicos também interfere na transformação do papel do professor como mediador na construção do conhecimento por meio do uso da tecnologia educacional.

4. ALGUNS RESULTADOS DA PESQUISA

Antes de iniciar com a coleta de dados específica, foram realizados pequenos experimentos em TCCs de acadêmicos do Curso de Administração da Faculdade da Serra Gaúcha.

Pode-se considerar que o processo estatístico do Sobek primeiro divide o texto em palavras. Os conceitos que tem maior ocorrência no texto são destacados para serem apresentadas no grafo resultante. Um conjunto de termos chamado de 'Stop Words' é utilizado para remover preposições, artigos ou palavras que não agregam informação sobre o texto (palavras muito comuns, que aparecem em diferentes contextos e não indicam qualquer tipo de informação relevante). Conceitos podem ser palavras simples (como ocorre na maioria dos casos), mas também podem ser um conjunto de palavras que tem um significado associado, como por exemplo "meio ambiente". As ligações são realizadas buscando conceitos que sejam considerados frequentes (aparecem no grafo), que estão distantes menos que 5 palavras uma da outra e não há ponto separando eles. Os conceitos mais frequentes podem possuir até 7 *links* com outros conceitos e esse número máximo reduz proporcionalmente à frequência do conceito. Considerando que os *links* são bidirecionais, alguns conceitos no grafo podem ter mais de 7 conexões (as 7 dele e mais alguma de outro conceito para ele). As conexões apresentadas são as mais frequentes que o conceito tem.

É possível analisar na figura 1 como ocorreu a aplicação da ferramenta Sobek

Figura 1: Processo de uso da ferramenta Sobek para análise textual
Fonte: Autores

A submissão dos TCCs ao software Sobek, seguiu o processo apresentado anteriormente:

a) Importar o texto para o Sobek;
b) Gerar o Grafo completo do TCC;
c) Gerar o Grafo por capítulos;
d) Analisar o Grafo e o Nodo por capítulo;

- A partir das práticas efetivadas chegou-se a uma média de escolha dos conceitos por página, para ser selecionado no momento de geração do grafo, conforme uma testagem realizada no segundo semestre de 2013 com alguns trabalhos que foram submetidos ao Sobek:

Testagem para definir o número de conceitos na geração dos grafos do SOBEK										
Esta testagem ocorreu com textos da disciplina de Diagnóstico Organizacional com acadêmicos da ADM da FSG, com trabalhos que tinham em média de 30 a 40 páginas.										
Capítulos	**TCC1**	**TCC2**	**TCC3**	**TCC4**	**TCC5**	**TCC6**	**TCC7**	**TCC8**	**TCC9**	**TCC10**
Grafo Geral Inicial	60	60	50	40	30	30	30	30	30	30
Introdução: Média de 3 pág	3	4	5	7	3	3	3	3	3	3
Referencial Teórico: Média de 10 pág	15	12	8	11	10	10	10	10	10	10
Metodologia: Média 3 de pág	5	4	5	3	3	3	3	3	3	3
Análise e Resultados: Média de 8 pág	5	6	7	8	8	8	8	8	8	8
Considerações: Média de 2 pág	3	2	3	2	3	3	3	3	3	3
Grafo Geral Final	30	30	30	30	30	30	30	30	30	30
O que se considerou desta testagem, é que em média os conceitos são selecionados pelo número de páginas do texto que foi submetido ao Sobek.										

Quadro 1: Média para a seleção do número de conceitos selecionadas para análise dos textos.
Fonte: Autores

- Analisar o Nodo de cada grafo: Nesta etapa é possível perceber o número de vezes que os conceitos (palavras) se repetem no capítulo. É interessante identificar na perspectiva da coerência e da conectividade, se o tema principal mantém-se no decorrer dos grafos gerados em cada capítulo. Assim como, nesta etapa, pode-se verificar a coesão, ou seja, foi viável observar erros gramaticais ou ortográficos.

Com esta metodologia empregada obtiveram-se alguns resultados, os quais auxiliaram para a criação da metodologia citada anteriormente e seguem alguns grafos

gerados para exemplificar a aplicação do método de análise textual com o apoio da ferramenta Sobek.

A figura 2 apresenta o Grafo Completo do texto gerado pelo Sobek:

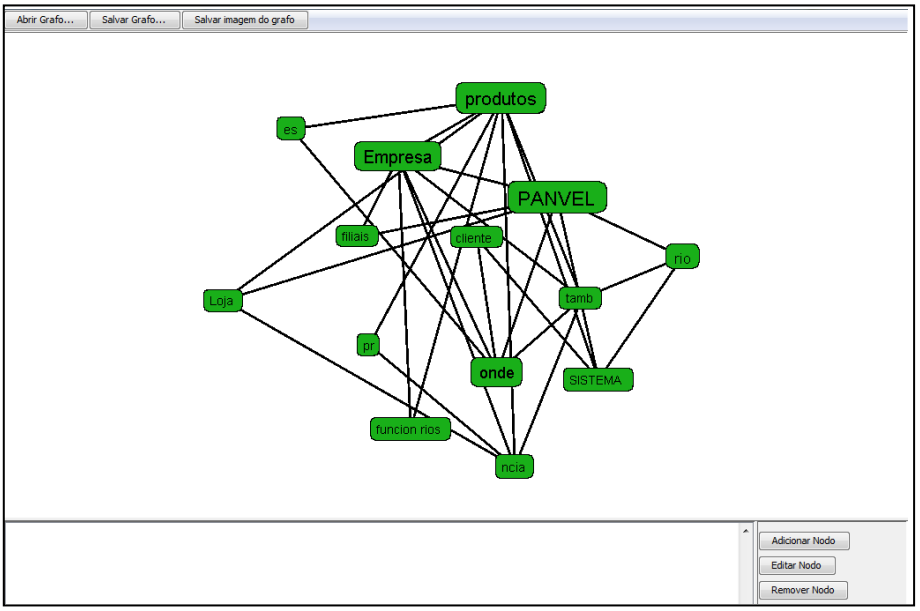

Figura 2: Grafo Geral do texto
Fonte: Autores

Em seguida para dar continuidade ao método de avaliação textual com apoio da ferramenta Sobek, foi gerado os Grafos de cada capítulo. Segue como exemplo a figura 3, que apresenta o grafo do capítulo da Fundamentação Teórica.

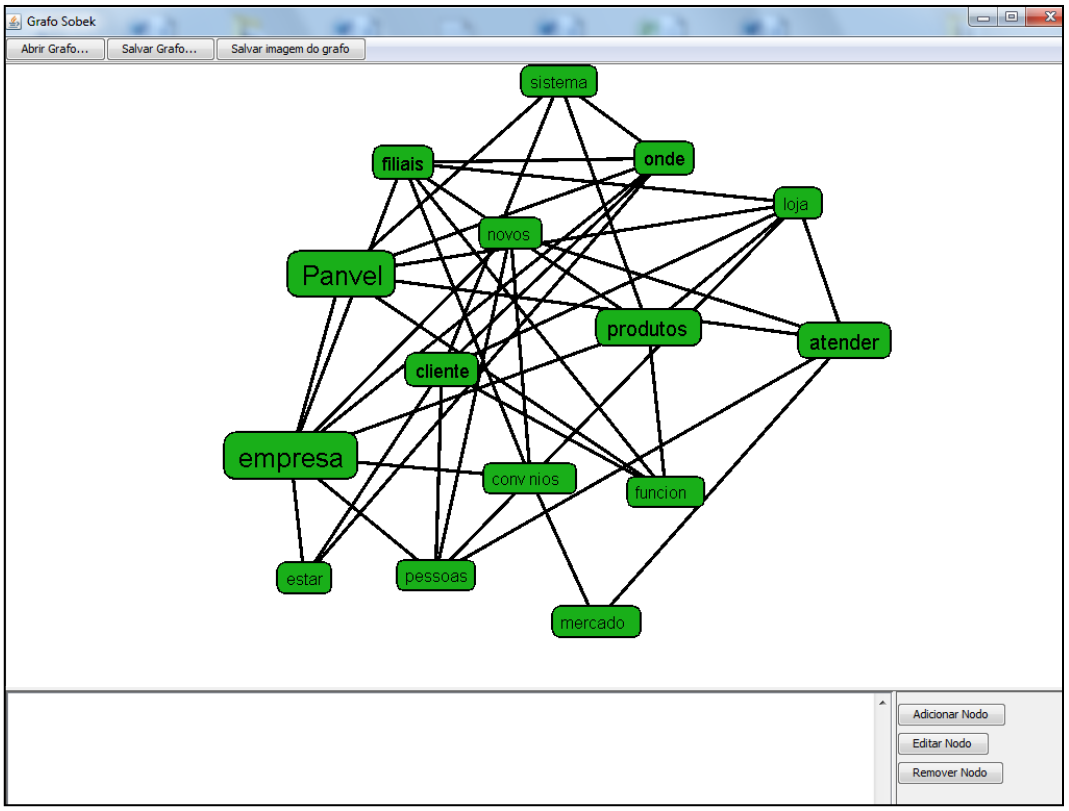

Figura 3: Grafo da Fundamentação Teórica
Fonte: Autores

Analisando os dois grafos, o Grafo Geral do texto e o Grafo da Fundamentação Teórica, podemos perceber que o objetivo do trabalho relata a continuidade, pois o propósito era apresenta um diagnóstico da empresa. Com isso é possível perceber que o trabalho responde ao critério de coerência e conectividade. No entanto, podem-se identificar os erros de coesão.

Por exemplo, com esses dois grafos simples, pode-se analisar um erro de coerência de forma claro. Isso visualizado na palavra "Onde" que consta no Grafo, exemplo "A Panvel busca um atendimento diferenciado, onde procura fazer com que o cliente se sinta seguro [...]", a palavra onde é utilizada de modo incorreto. Esta palavra é utilizada quando se refere a lugar e para ela aparecer no grafo, significa que consta muitas vezes no texto. Por isso, é sempre interessante que se analise o Grafo e o Nodo, pois o nodo mostra em que lugares do texto a palavra aparece.

A figura 4 apresenta o Grafo e o Nodo que identifica o erro de coesão.

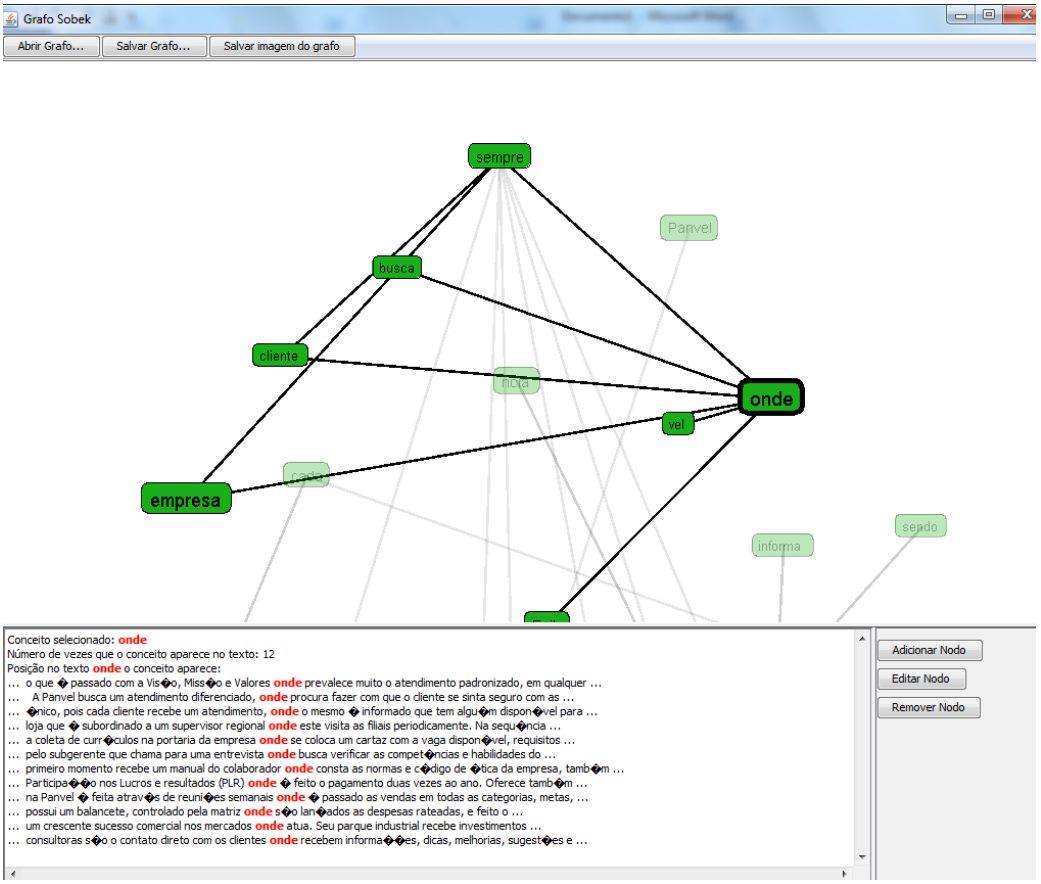

Figura 4: Análise de coesão
Fonte: Autores

E nessa perspectiva é realizada a investigação, analisando o grafo geral com os grafos por capítulos para identificar a coerência, a coesão e a conectividade de cada texto. Depois é entregue ao aluno um relatório de análise do seu texto com os respectivos grafos e nodos. Também foi proposto uma oficina para que o próprio aluno aprenda a submeter e analisar as suas produções escritas com o uso do Sobek

5. CONSIDERAÇÕES FINAIS

Esta pesquisa foi proposta para testar uma metodologia de aplicação da ferramenta Sobek no uso com análise de trabalhos acadêmicos na Faculdade da Serra

Gaúcha. Ela se refere a uma parte da investigação da tese de doutorado de uma das autoras pesquisadoras envolvida.

A partir desta primeira investigação realizada, todos os semestres seguintes estão sendo testados e registrados os experimentos com em média 20 TCCs por semestre para que se possa verificar a viabilidade de uso de uma ferramenta de mineração textual como apoio à escrita acadêmica, na busca pela qualidade e para responder os critérios de cientificidades na perspectiva da coerência, coesão e conectividade.

REFERÊNCIAS BIBLIOGRÁFICAS

AUSUBEL, David P. **A aquisição e retenção de conhecimentos:** uma perspectiva cognitiva. Lisboa, 2003.

BAKHTIN, M. **Estética da criação verbal.** Tradução: Maria Ermantina G. Gomes. São Paulo: Martins Fontes, 2000.

BARDIN, Laurence. **Análise de conteúdo.** Trad. Luís Antero Reto e Augusto Pinheiro. Lisboa: Edições 70, 2002.

BEAUGRANDE & DRESSLER (1983), **Texto e textualidade**. In: VAL, Maria da Graça C. Redação e textualidade. São Paulo: Martins Fontes, 1993.

FELDMAN, R.; SANGER, J. **The text mining handbook:** Advanced Approaches in Analyzing Unstructured Data. Cambridge, MA: Cambridge University Press, 2007.

GARCIA, Othon. **Comunicação em prosa moderna:** aprender a escrever, aprendendo a pensar. 22 ed. Rio de Janeiro: Editora FGC, 2002.

INFOESCOLA. **Coesão e Coerência**. Disponível em <www.infoescola.com/redacao/coesao-e-coerencia-textual>. Acesso em: 25 de jan. 2015.

LÉVY, P. **A máquina do universo** – criação, cognição e cultura informática. Porto Alegre: Artmed, 1998.

KENSKI, Vani M. **Educação e tecnologias:** o novo ritmo da informação. Campinas, SP: Papirus, 2007.

MOTTA-ROTH, D.; HENDGES, G. R. **Produção textual na universidade.** São Paulo: Parábola Editorial, 2010.

REATEGUI, E. **Sobek: a text mining tool for educacional applications**. In: INTERNATIONAL CONFERENCE ON DATA MINING, Las Vegas [s.n.], 2011. P.59-64.

RIOLFI, C.; IGREJA, S. G. **Ensinar a escrever no ensino médio:** cadê a dissertação? In: Educação e Pesquisa. São Paulo, v.36, n.1, p. 311-334. 2010.

SCHENKER, A. **Graph-Theoretic techniques for web content mining**. Florida: University of South Florida, 2003. PhD Thesis.

PIAGET, J. **A linguagem e o pensamento da criança.** Martins Fontes, São Paulo, 1986.

VYGOTSKY, L.S. **Pensamento e linguagem.** São Paulo: Martins Fontes,1987.

WARSCHAUER, M. and MESKILL, C. Technology and Second Language Teaching and Learning. In J. Rosenthal (Org.) **Handbook of Undergraduate Second Language Education**. Mahwah, NJ: Lawrence Erlbaum, 2000.