COLETA NA REDE SOCIAL TWITTER: UM *CORPUS* DO *ENEM 2016*

Marcus Vinicius Carvalho Guelpeli [1], Leila Maria Silva [2]

Abstract. *In the last years there has been a tendency of qualitative studies with the objective to understand the symbolic interactions between the individuals, being necessary to the researcher to select data representative of these processes. Corpus linguistics presents factors that can certainly help researchers to obtain and organize information to create their own text base to assist in the Text Mining process. It is intended to present the process of constructing a corpus with the theme Enem 2016 and to present the research methodology used to construct and organize a corpus. Based on the difficulties encountered in the area, the purpose of this research is to assist researchers, students and professionals in the construction of their own text - corpus base that may be useful in certain studies and specific research.*

Keywords: *Corpus; Linguistics of Corpus; Text Mining; Twitter.*

Resumo. *Nos últimos anos tem havido uma tendência de estudos qualitativos com o objetivo de compreender as interações simbólicas entre os indivíduos, sendo necessário ao pesquisador selecionar dados representativos desses processos. A linguística de corpus apresenta fatores que podem certamente auxiliar pesquisadores a obter e organizar informações para criar sua própria base de textos que auxiliem no processo de Mineração de Textos. Pretende-se apresentar o processo de construção de um corpus com o tema Enem 2016 e apresentar a metodologia de pesquisa utilizada para construir e organizar um corpus. A partir das dificuldades encontradas na área, a proposta desta pesquisa é auxiliar, pesquisadores, estudantes e profissionais na construção de sua própria base de textos - corpus que possam ser úteis dentro de determinados estudos e pesquisas específicas.*

Palavras Chave: *Corpus; Linguística de Corpus; Mineração de Textos; Twitter.*

[1] Professor na Universidades Federal dos Vales do Jequitinhonha e Mucuri (UFVJM) Diamantina – MG – Brasil. Email: marcus.guelpeli@ufvjm.edu.br
[2] Graduada em Sistemas de Informação – Universidades Federal dos Vales do Jequitinhonha e Mucuri (UFVJM) Diamantina – MG – Brasil. Email: leila-datas@hotmail.com

1 INTRODUÇÃO

O termo c*orpus* (plural: corpora) representa um conjunto de dados lingüísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por computador (SARDINHA, 2004).

"A partir da década de 1990, os *corpora* passam a ter papel fundamental nas pesquisas lingüísticas, pois data dessa época o início das contribuições advindas da Computação e da Lingüística Computacional" (ALUÍSIO e ALMEIDA, 2006).

"Assim, por meio de *corpus*, podem-se observar aspectos morfológicos, sintáticos, semânticos, discursivos, entre outros, bastante relevantes para uma pesquisa lingüística. Podem-se ainda descobrir fatos novos na língua, não perceptíveis pela intuição" (SARDINHA, 2000).

Este trabalho apresenta a construção de um *corpus*, composto por *tweets* sobre o tema Exame Nacional do Ensino Médio 2016 – Enem 2016, coletados da rede social *Twitter* durante o seu processo de realização no ano de 2016. Com a popularização do uso de redes sociais *online*, surgiu a oportunidade de estudos nesse tema com o uso de bases de dados, coletadas da rede social *Twitter*.

Segundo Benevenuto et al. (2011) "redes sociais *online* permitem o registro em larga escala de diversos aspectos da natureza humana relacionados à comunicação, à interação entre as pessoas e ao comportamento humano". Em geral, elas permitem que as pessoas interajam mais, mantenham contato com amigos e conhecidos e se expressem e sejam ouvidas por uma audiência local ou até mesmo global.

A rede social *Twitte*r foi utilizada como fonte de dados textuais deste trabalho por ser uma das redes sociais mais utilizadas no Brasil e por disponibilizar seu conteúdo, enquanto outras mantêm seus dados privados. Segundo Gomide (2012) "o *Twitter* fornece recursos para tornar seu conteúdo disponível. Através desses recursos pode-se obter a rede de seguidores das pessoas, as mensagens publicadas (*tweets*) por usuários, por região geográfica, por data ou até mesmo por palavras específicas".

A abordagem tradicional de coletar informações sobre temas comentados nas redes sociais é a de perguntar aos próprios membros do grupo como são as relações entres eles. Entretanto esse método é demorado e está sujeito a ter uma grande quantidade de abstenções. Além disso, vários estudos sobre a precisão dessas redes sociais, coletadas por meio de

questionários, mostram que "relatos individuais sobre interações sociais diferem substancialmente do que é realmente observado" (LANGE et al. 2004). "Deste modo, os pesquisadores da área têm se voltado para métodos automáticos para a coleta de dados em redes sociais" (MATSUO, HAMASAKI, *et al.*, 2006). Nas coletas dos *tweets* deste trabalho, foram utilizados *softwares* com o intuito de coletar automaticamente os *tweets*.

Segundo Aranha (2006) coleta é a etapa que tem como objetivo formar uma base de dados textual, conhecida na literatura como *Corpus* ou *Corpora*. Pode se dar de várias maneiras, porém todas necessitam de grande esforço, a fim de se conseguir material de qualidade e que sirva de matéria-prima para a aquisição de conhecimento.

Portanto, a etapa de recuperação e coleta de dados tem como função a criação de uma base de dados textual chamada *corpus* ou *corpora* que servirá de base para aplicar as técnicas de Mineração de Textos.

A formação do *corpus* sobre o Enem 2016 subsidiará pesquisas no descobrimento de informações úteis sobre o tema, tendo em vista que o Enem é um evento de grande relevância no Brasil e interesse social, e também cercado de polêmicas devido a vários problemas que vêm acontecendo nas edições anteriores. Problemas que merecem ser estudados em mais detalhes para que os responsáveis pelas políticas educacionais possam delinear estratégias de aprimoramento do exame.

Criado em 1998 com o objetivo de revolucionar a maneira como os estudantes do País são avaliados, o Exame Nacional do Ensino Médio (Enem) aumentou, ano a ano, de tamanho e de funções até alcançar uma das principais portas de entrada para o ensino superior no Brasil. Cresceu tanto que virou um problema: como ministrar uma prova em perfeitas condições de segurança, sigilo e baixo índice de erros para milhões de alunos em todo o território nacional?

As edições anteriores do Enem expuseram falhas que causam temor e descrença entre os estudantes e que prejudicam a reputação do Ministério da Educação. Problemas diversos como o furto das provas ocorrido em 2009, distribuição de provas com questões repetidas e outras faltando, divulgação do gabarito com erros, etc. Sendo o Enem, um exame de grande relevância, é interessante obter meios de conhecer melhor todo o seu processo e criar maneiras com o intuito de amenizar os problemas.

O fato da possível existência de conhecimento, até então, desconhecido, presente nestes textos é a contribuição para a área educacional. Com isso busca-se lançar um olhar sobre estas redes virtuais como espaços de construção e produção de discursos, manifestação das múltiplas "verdades" sociais e suas representações. Assim, esperamos neste trabalho, através

da descoberta de conhecimento sobre o Enem 2016, oferecer subsídios para a criação e/ou fortalecimento de iniciativas que visem melhorias em todo o processo do maior exame de avaliação do Brasil.

O texto está composto por quatro seções - seção dois define o termo *corpus*; seção três apresenta a metodologia aplicada no projeto, construção, organização e classificação do *corpus;* e quarta seção expõe as conclusões.

2 CORPUS

Segundo Sardinha (2004), "a Linguística de *Corpus* apropria-se da coleta e exploração de dados extraídos dos *corpora* para serem processados para fins de conhecimento de especificidades sobre o léxico de uma determinada língua ou variedade linguística". Ou seja, a Linguística de *Corpus* é um ramo da Linguística Aplicada que estuda os fenômenos linguísticos por meio de um *corpus*.

Para Sardinha (2004), Baker (2005), Sinclair (1995) e Tognini-Bonelli (2001) "*corpora* são coletâneas de textos selecionados e reunidos segundo critérios específicos, armazenados em formato eletrônico para que possam ser utilizados em análises linguísticas, representativos de uma língua ou variedade linguística e autênticos".

A linguística de *corpus* exerce grande influência na pesquisa linguística. Centros de pesquisas desenvolvidos de países como Inglaterra, País de Gales, Escócia, Noruega, Suécia e Dinamarca, dedicam-se à pesquisa baseada em *corpus* para a criação de materiais de apoio para trabalhos em diversas áreas. No Brasil, a linguística de *corpus* está em estágio inicial (SARDINHA, 2004).

De acordo com Aluísio e Almeida (2006), percebe-se que o surgimento do computador interferiu diretamente na criação, forma de armazenamento e exploração de um *corpus*, já que os recursos oferecidos pelo computador permitiram que uma quantidade muito grande de textos pudesse ser processada na tela em questão de segundos, viabilizando o teste de hipóteses de forma rápida e eficiente.

3 METODOLOGIA

A presente metodologia tem como objetivo estabelecer o processo desenvolvido ao longo dessa pesquisa. Os percursos metodológicos serão apresentados em quatro etapas: (i)

Projeto do *Corpus*; (ii) Construção do *Corpus*; (iii) Processamento do *Corpus* e (iv) Estatística do *Corpus*.

O Projeto do *Corpus* é o que também chamamos de desenho ou descrição do *corpus*. Trata-se da seleção do material, sendo importante destacar quais textos irão fazer parte do projeto, qual a natureza dos *corpora*, seus modos, sua temporariedade, entre outros aspectos essenciais para que se identifique a tipologia dos *corpora*.

O *corpus* de estudo aqui referido é composto por *tweets* coletados da rede social *Twitter* sobre o tema Enem no ano de 2016. Os *tweets* foram coletados de acordo com o cronograma do Enem 2016.

A Tabela 1 apresenta o cronograma de atividades sobre o Enem no ano de 2016:

Tabela 1: Cronograma do Enem 2016

Cronograma Enem 2016	
Inscrições do Enem	9 a 20 de Maio
1º Simulado Online	30 de Abril
2º Simulado Online	25 de Junho
3º Simulado Online	03 a 11 de Setembro
4º simulado Online	08 a 23 de Outubro
1ª Aplicação das Provas	5 e 6 de Novembro
2ª Aplicação das Provas	3 e 4 de Dezembro
Aplicação Prova para Pessoas Privadas de Liberdade	13 e 14 de Dezembro
Resultado do Enem	18 de Janeiro de 2017

A segunda parte do processo consiste na construção do *corpus*. Nesta fase foi realizada a coleta dos dados na rede social *Twitter* sobre o Enem 2016 utilizando três ferramentas de coleta: *Your Twapper Keeper*[1], *Collect Convert*[2] *e SherlockTM*[3].

[1] Vide http://twapperkeeper.com/index.html
[2] Vide https://github.com/ufeslabic
[3] Vide www.mtplnam.com.br

O processo de coleta foi iniciado no período de inscrições do Enem 2016 (09 de Maio de 2016) e encerrado no dia de divulgação dos resultados (18 de janeiro de 2017), para que assim, fosse feita uma análise dos acontecimentos relacionados ao evento durante todo o seu período de realização.

Não foram realizadas coletas nos meses de abril e junho de 2016, quando aconteceram o 1º e o 2º simulados *online*. Isso ocorreu devido à plataforma Hora do Enem (que realizou os simulados), ser uma ferramenta inédita, e ainda não conhecida. Neste período ainda estavam sendo levantadas informações a respeito da mesma e só mais adiante dos estudos identificamos a necessidade de realizar coletas nos períodos de aplicação dos simulados. Assim, dos quatro simulados *online* realizados, as coletas realizadas foram referente ao 3º e 4º simulados.

A primeira ferramenta utilizada foi o *Your Twapper Keeper* que possibilita arquivar dados do *Twitter* em tempo real diretamente no servidor de destino, para compor a base de documentos a ser trabalhada na etapa seguinte. É uma ferramenta que permite aos usuários arquivar, organizar e analisar os *tweets* com base em #hashtags, perfis ou palavras específicas. Falcão (2014) mostra que "o *Your Twapper Keeper* é um *software* utilizado em servidores do computador para captura e armazenamento de dados da plataforma".

O *Your Twapper Keeper* rastreia os *tweets* associados a uma determinada pesquisa, conforme os dados disponibilizados pelo usuário, para em seguida serem compilados em um arquivo geral, que pode ser de diversas extensões, como .csv, .html, entre outros. Porém a ferramenta apresenta a desvantagem de coletar os dados somente em tempo real e quando ocorre queda de energia ou de internet, a coleta é interrompida e os *tweets* são perdidos.

Assim, passou-se a realizar as coletas com outra ferramenta: o *collect convert*. Um *Script* de coleta automatizada de texto e imagem. Funciona a partir do acesso ao código fonte de uma página *web* que está na linguagem HTML (HyperText Markup Language, que significa Linguagem de Marcação de Hipertexto, é uma linguagem de programação utilizada na construção de páginas *web*) e consegue coletar *tweets* de até 7 dias anteriores.

Ao final do período das coletas foi realizada mais uma transição de ferramenta de coleta de *tweets*. As coletas passaram a ser realizadas pela *SherlockTM* (*Sherlock Text Mining*). É uma ferramenta desenvolvida em linguagem Java (Linguagem de programação interpretada orientada a objetos), destinado à área de mineração de textos. A *SherlockTM* integra um grupo

de funções como coleta automatizada de *tweets*, limpeza, organização e conversão dos arquivos para diferentes formatos.

A transição para o *SherlockTM* se justifica por ser uma ferramenta mais completa, pois além de realizar a coleta dos *tweets*, também possui funções auxiliares que ajudam no tratamento das informações, disponibilizando opções de pré-processamento de texto e de organização, mantendo as informações em diretórios organizados hierarquicamente.

Com o *SherlockTM* além de desonerar o tempo de pré-processamento de texto, um problema com custo humano alto, houve melhoria na qualidade do resultado uma vez que o *software* elimina os problemas relativos a possíveis erros humanos aliado à redução do tempo de execução das tarefas, sendo um diferencial em relação as outras ferramentas. Consegue fornecer bons resultados de forma consistente e contínua aumentando a qualidade do processo comparado com as outras ferramentas usadas anteriormente.

As transições de ferramentas durante o processo de coleta de *tweets* ocorreram na tentativa de tornar o processo mais eficaz e para obter maior qualidade no resultado final.

As coletas foram separadas em pastas de acordo com o mês correspondente. Para cada mês temos os termos das coletas que foram realizadas. Dentro de cada mês, foi realizadas coletas sobre o termo geral (palavra chave) “Enem 2016” durante todos os dias e nas datas específicas de eventos do Enem foi realizada a coleta sobre o evento. O *corpus* foi dividido em subpastas, conforme exemplificam as Figuras 1 e 2.

Figura 1: Pastas das Coletas dos *Tweets*

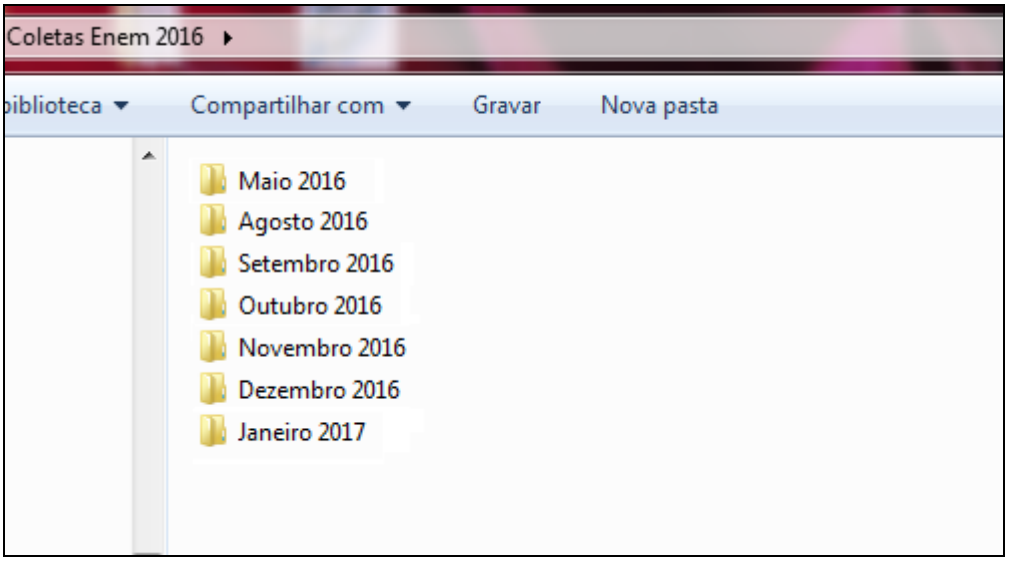
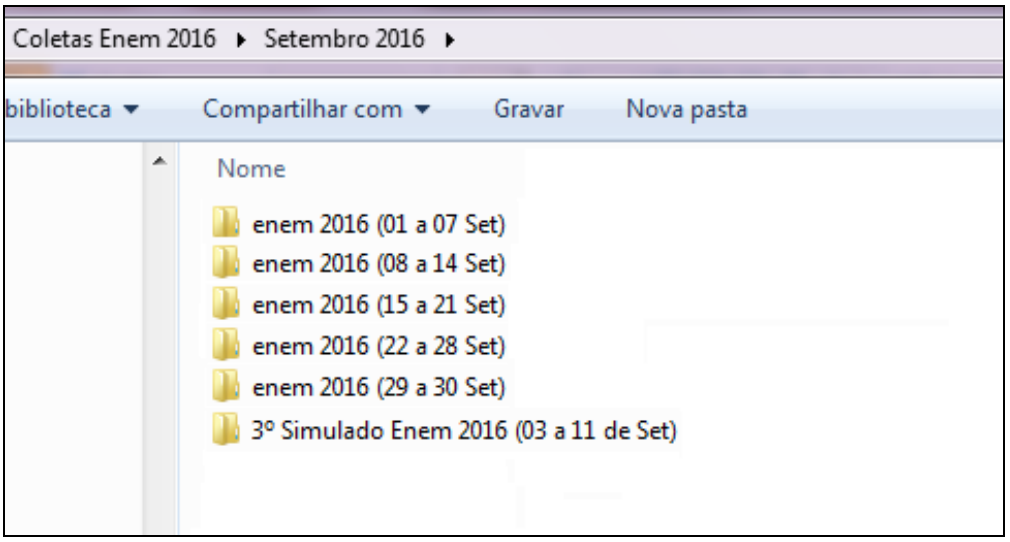


Figura 2: Subpastas com os termos (palavra chave) e as datas das coletas



Após a etapa de construção do *corpus*, o passo seguinte foi o processamento do *corpus* por meio da ferramenta computacional *SherlockTM*, utilizado para a "limpeza" dos textos, organização e conversão dos arquivos. Normalmente um *tweet*, contém algumas informações como: *id, texto, número de vezes que foi curtido, coordenadas,* dentre outros. Assim, ocorreu a limpeza e a formatação do *corpus* para o processamento computacional.

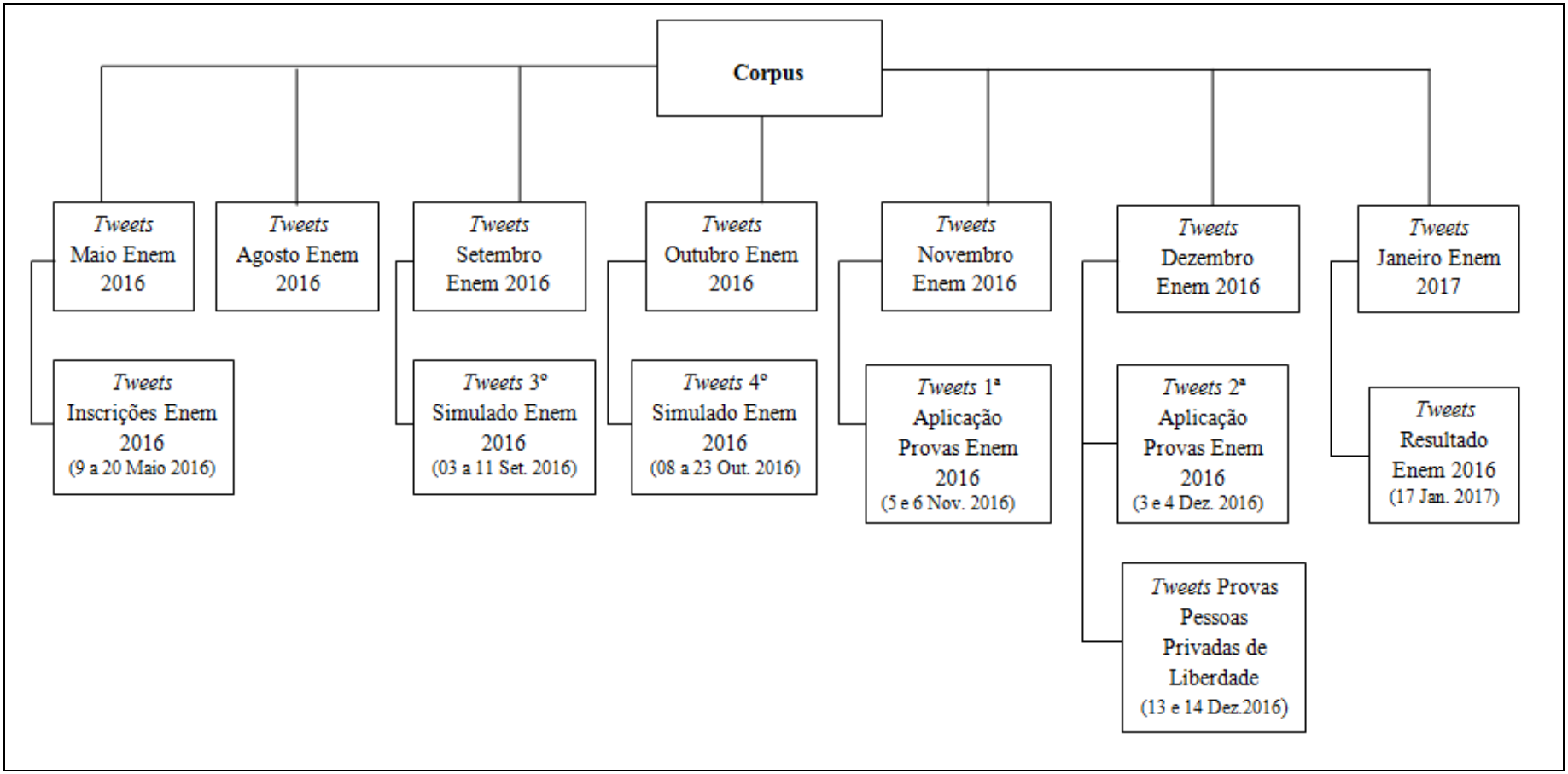
O *SherlockTM*, eliminou conteúdos considerados irrelevantes para o processo de Mineração do Texto, como *links*, acentos e *retweets*. Esses conteúdos foram excluídos para fornecer maior consistência nos resultados.

Com a finalidade de organizar e padronizar os arquivos, o *SherlockTM* também nomeou os *tweets* de acordo com as pastas em que estão inseridos, separados e colocados numa numeração sequencial. Esse padrão é mostrado na Figura 3.

Figura 3: Arquivos organizados na pasta após ser organizado pelo *SherlockTM*

Na Figura 4, é apresentada a estrutura organizacional dos *tweets* que compõem o *corpus* formado para este trabalho. A descrição dos períodos e datas das coletas foi realizada para permitir melhor compreensão do problema em análise. Todos os textos pertencem ao domínio Enem 2016 conforme descrito na etapa de projeto.

Figura 4: Diagrama do *corpus* construído neste trabalho



Na quarta etapa foi realizada a Estatística do *Corpus* para determinar a extensão do *corpus*. A análise de dados textuais, ou análise lexical, propõe a análise quantitativa de dados textuais. Torna-se possível, a partir da análise textual, descrever um determinado material. O uso *softwares* específicos para análise de dados textuais tem sido cada vez mais presente em pesquisas, especialmente naqueles estudos em que o *corpus* a ser analisado é bastante volumoso (Chartier & Meunier, 2011; Lahlou, 2012). Portanto, são usados programas de análise lexical para efetuar operações contagem e porcentagem de dados textuais como mostra a Tabela 2.

Tabela 2: Análise Estatística dos *Tweets* Enem 2016, composta por 2 arquivos: *Corpus* de *Tweets* Integral e *Corpus* de *Tweets* Pré-processado.

Arquivos	Caracteres	Caracteres e Espaços	Palavras	Palavras e Números	Porcentagem de Números	Frases	Porcentagem de Frases Repetidas
***Corpus* Integral**	9.076.846	10.650.837	1.567.559	1.891.997	17,15%	151.842	58,34%
***Corpus* Pré-processado**	6.817.768	8.183.776	1.210.742	1.497.214	19,13%	98.614	50,33%
Diferença	2.259.078	2.467.061	356.817	394.783	-	53.228	-

A Tabela 2 mostra a estatística do *corpus*, referente a duas análises. A primeira é referente ao *corpus* de *tweets* integral, ou seja, completa como os *tweets* foram coletados e a outra para o *corpus* de *tweets* pré-processados, ou seja, após a limpeza e manipulação. As linhas têm valores relacionados à quantidade de Caracteres (símbolo, letra ou número), Caracteres e Espaços, Palavras, Palavras e Números, Porcentagem de Números, Frases e Porcentagem de Frases Repetidas, e as colunas os valores correspondentes ao *Corpus* Integral, ao *Corpus* Pré-processado e por fim a diferença entre os dois. É importante observar que houve uma redução do *corpus* integral em relação ao *corpus* pré-processado, proporcionando um ganho qualitativo e quantitativo para o processo computacional. No *corpus* Integral há um texto com 9.076.846 caracteres, já no *corpus* Pré-processado são 6.817.768 caracteres, havendo uma diferença de 2.259.078 caracteres. As estatísticas do *corpus* foram calculadas com a utilização do *software Fine Count*[4].

[4]Vide http://www.tilti.com/software-for-translators/finecount/

4 CONCLUSÃO

Este artigo apresentou técnicas de desenvolvimento da Linguística de *Corpus*, mostrando os procedimentos para projetar, construir, processar e realizar a estatística de um *corpus* a partir da rede social *Twitter* aplicadas a fim de descobrir informações de tendências e padrões sobre o tema Enem 2016.

Este *corpus* tem a finalidade de viabilizar estudos da influência destes textos dentro do tema em questão através das técnicas de Mineração de Textos. "Com as técnicas de Mineração Textual não apenas pode ser realizada a compreensão dos textos, mas também trilhar os passos até a avaliação e implementação dos mesmos" (CHAPMAN, CLINTON, *et al.*, 2000). A técnica é utilizada para descobrir, de forma automática, informações (padrões e anomalias) em textos.

Então torna-se viável a utilização da Mineração de textos para analisar o conteúdo gerado pelos usuários na rede social *Twitter* sobre o tema Enem 2016, possibilitando descobrir informação desconhecida e relevante que pode auxiliar no aprimoramento do Enem.

A próxima etapa deste trabalho é obter o relacionamento entre os documentos textuais, verificando o grau de similaridade e a formação de grupos naturais, então a tarefa a ser escolhida é a Mineração de Texto com C*lusterização*. Assim, este trabalho propõe a utilização do modelo Cassiopeia, que conforme Guelpeli (2012), mostrou ser uma solução viável para minerar os *tweets* e aplicar o algoritmo clusterizador gerando *clusters* ou grupos de textos que serão analisados para gerar conhecimento.

O *corpus* construído neste trabalho faz parte dos estudos desenvolvidos no grupo de pesquisa Mineração de Textos e Processamento de Linguagem Natural e Aprendizado de Máquina (MTPLNAM), que também já produziu outros corpora, conforme os trabalhos de Oliveira e Guelpeli (2014), Fernandes e Guelpeli (2014) e Guelpeli e Fernandes (2016).

O grupo MTPLNAM tem como objetivo pesquisar, gerar conhecimentos e desenvolver aplicações sobre mineração de texto (MT), processamento de linguagem natural (PLN) e aprendizagem de máquina (AM), sua página pode ser acessada em http://www.mtplnam.com.br, onde estão outros corpora produzidos pelo grupo e disponibilizados para a comunidade científica para testes e trabalhos.

Espera-se, através de trabalhos futuros que serão embasados no trabalho atual, contribuir na oferta de subsídios para a criação e/ou fortalecimento de iniciativas que visem melhorias

no processo de realização do Enem, bem como auxiliar, por exemplo, os gestores públicos na criação/consolidação de ações afirmativas sobre o exame.

REFERENCIAS

Aluísio, S. M., & Almeida, B. G. M. (2006). O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio*, *4*(3), 156-178.

Aranha, C.; Passos, E. (2006) A Tecnologia de Mineração de Textos. Lab. ICA Elétrica PUC-Rio. RESI-Revista Elerônica de Sistemas de Informação, Nº2.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, *7*(2), 223-243.

Benevenuto, F., Almeida, J. M., & Silva, A. S. (2011). Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. *Porto Alegre: Sociedade Brasileira de Computação*.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chartier, J. F., & Meunier, J. G. (2011). Text mining methods for social representation analysis in Large Corpora. *Papers on Social Representations*, *20*(37), 1-47.

Falcão, de S. P. A Genealogia das Lutas Multitudinárias em Rede. O #vemprarua no Brasil. Mestrado, Universidade Federal do Rio de Janeiro, 2014.

Fernandes, H M; Guelpeli, M. V. C. Creación de corpus en lengua española para suutilización en testes acerca de Sumarización Automática. In: 6th International Conference on Corpus Linguistics-CILC 2014, 2014, Las Palmas de Gran Canaria. 6th International Conference on Corpus Linguistics-CILC 2014, 2014.

Gomide, J. S. (2012). Mineração de redes sociais para detecção e previsão de eventos reais. — Belo Horizonte, 85 f.

GUELPELI, M.V.C; Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização. 2012. 220f. Tese (Doutorado em Computação) – Universidade Federal Fluminense, Rio de Janeiro, 2012.

Guelpeli, M. V. C.; Fernandes, H. M. Input a Word, Analyze the World: Selected Approaches to Corpus Linguistics... ISBN-13: 978-1-4438-8513-3 e ISBN-10: 1-4438-8513-4 1ª. ed. United Kingdom: Cambridge Scholars Publishing, 2016. v. I. 521p.

Lahlou, S. (2001). Text mining methods: an answer to Chartier and Meunier. *Papers on Social Representations*, *20*(38), 1-7.

Lange, D. D.; Agneessens, F., & Waege, H. (2004). Asking social network questions: a quality assessment of different measures. *Metodoloski zvezki*, *1*(2), 351.

Matsuo, Y., Hamasaki, M., Nakamura, Y., Nishimura, T., Hasida, K., Takeda, H., ... & Ishizuka, M. (2006, July). Spinning multiple social networks for semantic web. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, No. 2, p. 1381). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Oliveira, R.R.; Guelpeli, M.V.C. Building a Corpus in Italian Written Language. In: 6th International Conference on Corpus Linguistics (CILC2014). Las Palmas de Gran Canaria, Espanha, 2014. No prelo.

Sardinha, T. B. (2000). Lingüística de corpus: histórico e problemática. *Delta*, *16*(2), 323-367.

Sardinha, T. B. (2004). *Lingüística de corpus*. Editora Manole Ltda.

Sinclair, J. (1991). Corpus Concordance Collocation.Oxford: Oxford University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work* (Vol. 6). John Benjamins Publishing.