"*Corpus* Linguístico UFSM": Construção e etiquetagem

Débora Spanamberg Wink (UFSM)
Sabrina Damiani Schmidt (UFSM)
Sara Regina Scotta Cabral (UFSM)

RESUMO: Este trabalho tem como principal objetivo explicar o processo de construção e de organização do "*Corpus* Linguístico UFSM". Para tanto, faz-se uso dos estudos de Halliday (1994), Halliday e Matthiessen (1999, 2004, 2014) e seus seguidores para demonstrar a necessidade de um *corpus* bem estruturado e organizado para uma análise sistêmico-funcional. Além disso, traz-se Miller (2012) para justificar as escolhas dos gêneros que compõem o *corpus*. Para a metodologia deste trabalho, busca-se suporte nos estudos de Bick (1996) e de Berber Sardinha (2004) sobre Linguística de *Corpus* para o processo de etiquetagem morfossintática dos textos. Por fim, apresentam-se os resultados obtidos na coleta e na etiquetagem do "*Corpus* Linguístico UFSM", além da expectativa de compartilhamento *on-line* desse *corpus*.

PALAVRAS-CHAVE: Linguística Sistêmico-Funcional. Linguística de *Corpus*. *Corpus* Linguístico UFSM, Etiquetagem.

ABSTRACT: This work aims to explain the process of construction and organization of the *UFSM Linguistic Corpus*. We based our research on the studies of Halliday (1994), Halliday and Matthiessen (1999, 2004, 2014), and their followers to demonstrate the need for a well-structured and organized *corpus* in a systemic-functional analysis. Besides, we draw on Miller's (2012) work to explain the genres chosen to make up the *corpus*. Corpus Linguistics (BICK, 1996; BERBER SARDINHA, 2004) provides the methodology for the process of tagging texts. Finally, we present the results obtained in the collection and tagging of *UFSM Linguistic Corpus*. In addition, we intend to share this *corpus* online.

KEYWORDS: Systemic Functional Linguistics. *Corpus* Linguistics. UFSM Linguistic *Corpus*, tagging.

1 CONSIDERAÇÕES INICIAIS

O "*Corpus* Linguístico UFSM" foi construído com o objetivo de prover suporte a trabalhos a serem desenvolvidos por pesquisadores da área de Letras e por usuários interessados no trabalho com textos, especialmente os estudiosos da área de Linguística Sistêmico-Funcional. Para isso, o *corpus* em questão está constituído de gêneros de três áreas do discurso (midiático, político e acadêmico) que foram organizados conforme as orientações da Linguística de *Corpus* (BERBER SARDINHA, 2004). Ademais, o "*Corpus* Linguístico UFSM" é fruto do projeto "Mídia, Política e Gramática Sistêmico-Funcional", que parte da perspectiva hallidayana de que, mesmo inconscientemente, o usuário de uma língua faz escolhas para realizar a linguagem e, assim, participa de um sistema de opções que estão à sua

disposição. Dessa forma, ao se analisar cientificamente um texto, é possível sistematizar as escolhas do usuário da língua e ainda revelar como essas escolhas estão funcionalmente organizadas dentro da estrutura desse texto. Partindo dessa perspectiva, tanto o projeto quanto este trabalho tem como aporte teórico os estudos de Halliday (1994) e Halliday e Matthiessen (1999, 2004, 2014), bem como de alguns de seus seguidores (THOMPSON, 2004; BLOOR; BLOOR, 1995; MARTIN; MATTHIESSEN; PAINTER, 1997). Assim, o objetivo deste artigo é explicar o processo de construção, organização e etiquetagem do "*Corpus* Linguístico UFSM" e, para melhor apresentá-lo, este trabalho divide-se em duas grandes seções, além desta seção introdutória e da seção *Considerações finais*. A primeira grande seção (*Referencial teórico*) é dividida em três subseções: (1) *Linguística Sistêmico-Funcional*, (2) *Gêneros do discurso* e (3) *Linguística de Corpus*. Em contrapartida, o relato da construção e da etiquetagem do *corpus*, além de uma amostragem de análise, podem ser encontrados detalhadamente em outra grande seção deste trabalho, intitulada *Metodologia*. Por fim, têm-se as *Considerações Finais*.

2 REFERENCIAL TEÓRICO

2.1 Linguística Sistêmico-Funcional

A Linguística Sistêmico-Funcional adquiriu grande notoriedade a partir dos estudos de Michael Halliday. Essa teoria adota uma abordagem descritiva para explicar o funcionamento da linguagem em diferentes contextos de uso. Sob essa perspectiva, os falantes selecionam – ainda que inconscientemente – os elementos linguísticos mais adequados para o sucesso de uma determinada situação comunicativa, de modo que é possível identificar um sistema de escolhas a partir da descrição do uso real da língua. Sendo assim, o texto sempre é analisado levando em consideração as informações do contexto de situação e do contexto de cultura.

O contexto de cultura é mais amplo que o de situação. Ele se refere a características típicas de determinados grupos sociais ou de práticas sociais já institucionalizadas (exemplo: entrevistas de emprego), enquanto o contexto de situação é mais restrito a uma interação comunicativa específica (exemplo: "Pedro realiza entrevista para vaga de fisioterapeuta").

Cada interação, portanto, possui características situacionais particulares, que variam em três aspectos: no campo (a natureza e os objetivos da interação), nas relações (quem são os participantes e como é seu comportamento linguístico) e no modo (como a língua e outros sistemas semióticos são organizados). (HALLIDAY; MATTHIESSEN, 2014).

Além dessa variação contextual, outra característica intrínseca à linguagem é a sua funcionalidade (HALLIDAY; MATTHIESSEN, 2014, p. 41), e, para assinalar a distinção entre a função oracional em um contexto específico e as funções inerentes à língua, os autores preferem o termo metafunção. Nesse sentido, o texto pode assumir três metafunções básicas – ideacional, interpessoal e textual, conforme exemplifica o Quadro 1.

Quadro 1 - Integração multifuncional da oração

Metafunções	A denunciada	matou	seu filho recém-nascido	em 19.9.1997.
Ideacional	Participante	Processo	Participante	Circunstância
Interpessoal	Sujeito	Finito e Predicador	Resíduo	
Textual	Tema	Rema		

Fonte: Adaptado de Fuzer; Cabral 2014, p. 35.

É importante ressaltar que a oração não assume uma metafunção ou outra isoladamente; todas coocorrem simultaneamente. O que se altera é o enfoque e a metodologia da análise dos elementos gramaticais, já que, para cada metafunção, a oração é encarada de maneira distinta. Dessa forma, ao analisar a oração do Quadro 1, vê-se que, na perspectiva da metafunção ideacional, o texto é compreendido como uma representação das experiências humanas, isto é, tem um processo (material, mental, relacional, verbal, existencial e comportamental), participantes (Ator, Meta, Experienciador, etc.) e circunstâncias. Na metafunção interpessoal, em que o texto é uma troca entre dois (ou mais) participantes da interação, a oração apresenta Sujeito, Finito e Predicador e Resíduo. Já na textual, em que o texto assume a função de mensagem organizada, a oração é dividida em Tema e Rema.

Uma vez que, para analisar sistêmico-funcionalmente a língua, é preciso um texto falado ou escrito, tem-se a necessidade de se formar um *corpus*. Para isso, faz-se uma seleção de textos (subseção 2.3) baseando-a em uma perspectiva textual para justificar a (s) escolha (s) de gênero (s), como mostra a subseção a seguir.

2.2 Gêneros do discurso

Conforme lembra Miller (2012), "uma classificação une itens com base em algum conjunto de semelhanças" (p. 22), exatamente como acontece com o "*Corpus* Linguístico UFSM", isto é, o *corpus* é dividido em três classificações principais (Discurso Acadêmico, Discurso Midiático e Discurso Político) e, em cada uma delas, encontram-se textos com características semelhantes. Tais classificações utilizadas para organizar o *corpus* levam em conta outra afirmativa de Miller que diz que "a classificação do discurso será retoricamente sólida se contribui para uma compreensão de como o discurso funciona – isto é, se reflete a experiência retórica do povo que cria e interpreta o discurso". (2012, p. 22). Dessa forma, os discursos que compõem, por exemplo, a primeira classificação principal (Discurso Acadêmico – subseção 2.2.1) são textos de cunho acadêmico, isto é, artigos científicos, dissertações de Mestrado e teses de Doutorado. Da mesma maneira, a segunda classificação principal (Discurso Midiático – subseção 2.2.2) é composta por textos midiáticos como as notícias, e, por sua vez, a terceira principal classificação (Discurso Político – subseção 2.2.3) compõe-se de textos de cunho político como, por exemplo, declarações, pronunciamentos, palavras e discursos de políticos brasileiros.

Nas subseções a seguir, encontram-se a sustentação teórica para as escolhas de tais gêneros do discurso para a composição do "*Corpus* Linguístico UFSM"

2.2.1 Discurso Acadêmico

Segundo Halliday (2004, p. 125, 126), o discurso acadêmico e/ou científico é "tipicamente construído a partir de uma sequência de passos ligados entre si, de tal modo que a qualquer momento uma bateria inteira de passos anteriores pode ser codificada como motivo para a próxima"[1]. Ademais, ao se construir um discurso científico, duas condições semióticas devem ser levadas em consideração, como bem lembra Halliday (2004). Segundo o autor, a primeira delas é a condição técnica, ou seja, "a gramática tem que criar significados técnicos, fenômenos puramente virtuais que existem apenas no plano semiótico, como termos de uma

[1] Todas as traduções deste trabalho são responsabilidade das autoras.

teoria; não isoladamente, mas de forma organizada através de taxonomias elaboradas" (HALLIDAY, 2004, p. 123). A segunda condição semiótica que deve aparecer no texto acadêmico/científico é a racionalidade, isto é, "a gramática tem que criar uma forma de discurso para o raciocínio a partir da observação e da experiência e precisa também tirar conclusões gerais e progredir de um passo para outro em sequências de argumento lógico" (HALLIDAY, 2004, p. 123). Partindo dessa perspectiva, ao "*Corpus* Linguístico UFSM" interessa apenas os artigos publicados em periódicos da Universidade Federal de Santa Maria e de dissertações de Mestrado, bem como teses de Doutorado, defendidas na mesma instituição de ensino.

2.2.2 Discurso Midiático

Dentre as diversas áreas de estudo dos mecanismos sociais e culturais, várias pesquisas têm dado especial atenção aos fenômenos midiáticos, pois "eles são econômica e politicamente motivados, vinculados à evolução da ciência e tecnologia e, como a maioria dos domínios da vida humana, sua existência está intimamente ligada ao uso da linguagem". (SPITULNIK, 1993, p. 293). Ao "*Corpus* Linguístico UFSM", porém, interessa especialmente o discurso utilizado pelos profissionais do jornal (LAGE, 1993; RABAÇA; BARBOSA, 2001) nos mais diversos gêneros que compõem esse meio de comunicação (notícia, reportagem, anúncio, artigo de opinião, etc.).

2.2.3 Discurso Político

Desde Aristóteles, sabe-se que o homem é um ser político. E se, por muito tempo, foi a Retórica Clássica que se ocupou dos desenvolvimentos do discurso político, hoje se pode contar com outros instrumentos para o estudo dessa área. Os estudos em análise do discurso podem contribuir para um melhor entendimento do contexto de produção, da circulação e do consumo dessas peças, além de identificar mecanismos linguísticos produtivos na relação entre governantes e governados.

Com o decorrer do tempo, o discurso político midiatizou-se e espetacularizou-se. Em outras palavras, os mecanismos midiáticos têm incrementado as ações políticas e mudado as

estratégias de persuasão através de novas tecnologias, assim como os jornais. A partir disso, o "*Corpus* Linguístico UFSM" é composto por pronunciamentos, discursos e declarações à imprensa da ex-presidente Dilma Rousseff e deve ser expandido através da coleta de discursos de outros políticos brasileiros.

2.3 Linguística de *Corpus*

A Linguística de *Corpus* (LC), juntamente com o Processamento de Linguagem Natural (PLN), compõe a Linguística Computacional, estudada nacional e internacionalmente. A LC encarrega-se da compilação e da análise de *corpora* e, atualmente com o avanço tecnológico, tem se desenvolvido através do trabalho com *corpora* eletrônicos, uma vez que a internet tem se mostrado uma forte aliada na coleta e na análise dos mais variados *corpora*. No Brasil, a LC iniciou com os estudos de Berber Sardinha (2004), seguindo a linha de raciocínio de estudiosos estrangeiros como Biber, Conrad e Reppen (1998), além de Tognini-Bonelli (2001).

Segundo os estudiosos da LC, *corpora* (no singular, *corpus*) são as grandes compilações de textos em formato eletrônico ou não, orais ou escritos, sincrônicos ou diacrônicos, mais abrangentes ou mais específicos, variando de acordo com o objetivo de estudo. Além disso, no manuseio do *corpus*, LC faz uso do 'suíte' *WordSmith Tools* (disponível para *download* em várias versões). Como já sugere o nome "Tools", o *WordSmith Tools* é um *software* hospedeiro. Dentro dele, existem três programas (*WordList*, *Concordance* e *KeyWords*). O *WordList* é encarregado de revelar em forma de lista as palavras mais frequentes no *corpus* e o número dessas ocorrências. Por sua vez, o *Concordance* é a ferramenta que permite ao pesquisador visualizar as concordâncias das palavras mais frequentes, reveladas pelo *WordList*. Já o *KeyWords* é responsável por calcular as palavras-chave e também por extrair palavras de uma lista de frequências.

Além do uso das ferramentas presentes no *WordSmith Tools*, os pesquisadores podem utilizar ferramentas etiquetadoras morfossintática, sintática e/ou discursiva como, por exemplo, a *Visual Interactive Syntax Learning* (VISL), que é explorada na subseção 3.2 deste artigo.

A seguir, apresenta-se a metodologia de elaboração deste trabalho.

3 METODOLOGIA

3.1 Coleta

Baseando-se nos estudos dos grandes estudiosos da Linguística Sistêmico-Funcional – Halliday (1985; 1994), Halliday e Matthiessen (1999; 2004; 2014) e seus seguidores, este trabalho tem como principal objetivo descrever o processo de construção do "*Corpus* Linguístico UFSM", a ser disponibilizado na internet para propiciar a graduandos, pós-graduandos e docentes um material de uso real para o exame de práticas discursivas sob o viés da Linguística Sistêmico-Funcional. Para tanto, são utilizados exemplares de discursos acadêmico, jornalístico e político produzidos no Brasil.

Para a constituição do *corpus*, foram coletados, no decorrer dos anos de 2014, 2015, 2016 e 2017, textos das mais variadas fontes digitais, entre elas: Portal do Planalto (discursos políticos), site da Universidade Federal de Santa Maria (dissertações e teses do curso de Letras e artigos de periódicos), Jornal Folha de São Paulo, Jornal O Estadão, Portal de Notícias G1 e Jornal Zero Hora (notícias). Uma vez coletados e reunidos, em meio digital, os três tipos de discursos (acadêmico, midiático e político), de acordo com as orientações de Berber Sardinha (2004), os textos foram gravados em arquivos .txt para viabilizar o uso da ferramenta computacional *WordSmith Tools 7.0*. (SCOTT, 2016).

Além do arquivamento em .txt, os textos foram armazenados em mais dois formatos distintos, quando necessário: .doc e .pdf. Sucedendo a coleta do *corpus* e a armazenagem, fez-se a classificação dos três tipos de discursos (acadêmico, midiático e político), bem como sua separação no arquivamento. A Figura 1 ilustra a organização dos textos no *corpus*.

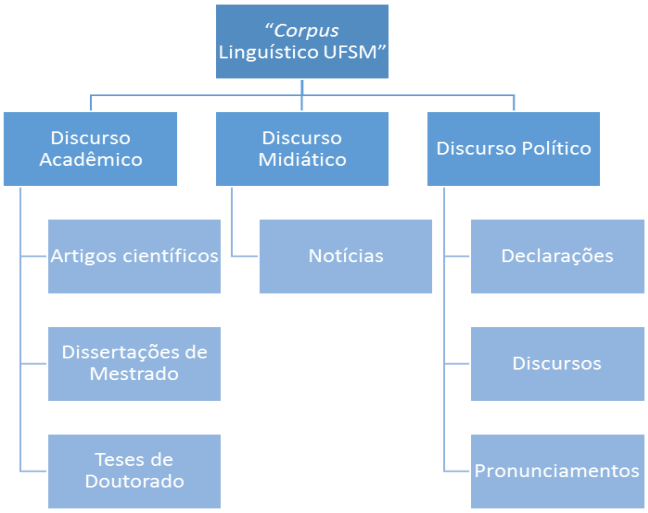

Figura 1 – Organização dos textos que compõem o "*Corpus* Linguístico UFSM"

A Figura 1 explicita a organização do "*Corpus* Linguístico UFSM" revelando que o *corpus* se divide em três classificações principais (Discurso Acadêmico, Discurso Midiático e Discurso Político). A primeira classificação principal - Discurso Acadêmico (DA) - é subdividida em três – artigos científicos, dissertações e teses. Tanto a subclassificação dissertações quanto a teses são divididas em dissertações de Linguística, dissertações de Literatura, teses de Linguística e teses de Literatura, sendo todas da Universidade Federal de Santa Maria (UFSM), bem como os artigos científicos que são coletados do Portal de Periódicos da UFSM. A segunda principal classificação é o Discurso Midiático (DM) que é composta, temporariamente, pela subclassificação Notícia, a qual ainda é dividida de acordo com seus assuntos/temas e, posteriormente, pela fonte, como pode ser visto na Figura 2.

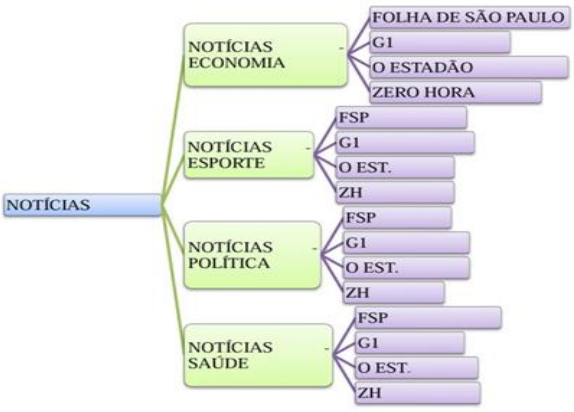

Figura 2 – Classificação principal Discurso Midiático e suas divisões

Como pode se observar, a Figura 2 ilustra a organização da subclassificação Notícias, parte da classificação principal Discurso Midiático. A subclassificação Notícias é dividida em Notícias de Economia, Notícias de Esporte, Notícias de Política e Notícias de Saúde e cada uma dessas subclassificações é dividida de acordo com a sua fonte (Folha de São Paulo, G1, O Estadão e Zero Hora). Após o DM, está a terceira e última principal classificação que é o Discurso Político (DP), o qual se subdivide em Declarações à imprensa, Discursos/ Palavras e Pronunciamentos.

Por possuírem nomes muitos extensos para suportarem suas classificações principais e subclassificações, cada arquivo possui um código que pode ser observado na Figura 3.

Legendas – arquivos *corpus*:

#**NFSPEC** – Notícias Folha de S. Paulo – Economia (formato Word)
%**NFSPEC** – Notícias Folha de São Paulo – Economia (formato txt)

#**NFSPESP** – Notícias Folha de S. Paulo – Esporte (formato Word)
%**NFSPESP** – Notícias Folha de São Paulo – Esporte (formato txt)

#**NFSPP** – Notícias Folha de São Paulo – Política (formato Word)
%**NFSPP**– Notícias Folha de São Paulo – Política (formato txt)

#**NFSPS** – Notícias Folha de São Paulo – Saúde (formato Word)
%**NFSPS** – Notícias Folha de São Paulo – Saúde (formato txt)

#**NG1EC** – Notícias G1 – Economia (formato Word)
%**NG1EC** – Notícias G1 – Economia (formato txt)

#**NG1ESP** – Notícias G1– Esporte (formato Word)
%**NG1ESP** – Notícias G1 – Esporte (formato txt)

#**NG1P** – Notícias G1 – Política (formato Word)
%**NG1P**– Notícias G1 – Política (formato txt)

#**NG1S** – Notícias G1 – Saúde (formato Word)
%**NG1S** – Notícias G1 – Saúde (formato txt)

#**NOEEC** – Notícias O Estadão – Economia (formato Word)
%**NOEEC** – Notícias O Estadão – Economia (formato txt)

#**NOEESP** – Notícias O Estadão – Esporte (formato Word)
%**NOEESP** – Notícias O Estadão – Esporte (formato txt)

#**NOEP** – Notícias O Estadão – Política (formato Word)
%**NOEP**– Notícias O Estadão – Política (formato txt)

#**NOES** – Notícias O Estadão – Saúde (formato Word)
%**NOES** – Notícias O Estadão – Saúde (formato txt)

#**NZHEC** – Notícias Zero Hora – Economia (formato Word)
%**NZHEC** – Notícias Zero Hora – Economia (formato txt)

#**NZHESP** – Notícias Zero Hora – Esporte (formato Word)
%**NZHESP** – Notícias Zero Hora – Esporte (formato txt)

#**NZHP** – Notícias Zero Hora – Política (formato Word)
%**NZHP**– Notícias Zero Hora – Política (formato txt)

#**NZHS** – Notícias Zero Hora – Saúde (formato Word)
%**NZHS** – Notícias Zero Hora – Saúde (formato txt)

Figura 3 – Legendas dos arquivos referentes ao *corpus* do Discurso Midiático ("Notícias"), de acordo com o formato de cada arquivo

A Figura 3 explicita o código dado a cada arquivo coletado, por exemplo N para notícia, FSP para Jornal Folha de São Paulo, EC para notícias de Economia, E para Jornal O Estadão, ES para notícias de Esporte, P para notícias de Política, e assim por diante. Dessa forma, quando, por exemplo, o arquivo é uma notícia de economia retirada do portal G1, o código passa a ser N (notícia), G1 (portal G1), EC (economia), ficando assim: NG1EC. Para distinguir os arquivos com a mesma codificação, mas com o formato diferente, deu-se os códigos # para formato *doc.*, % para formato *txt.* e, quando necessário, @ para PDF.

Uma vez explicado o processo de construção e de organização do "*Corpus* Linguístico UFSM", comentar-se-á, na subseção seguinte, sobre o processo de etiquetagem do *corpus*, bem como sobre o etiquetador utilizado.

3.2 *Visual Interactive Syntax Learning* (VISL)

A *Visual Interactive Syntax Learning* (VISL) é uma pesquisa e um projeto em desenvolvimento, desde 1996, pertencente ao Instituto de Linguagem e Comunicação (ISK), da Universidade do Sul da Dinamarca (SDU). O site da ferramenta VISL (http://beta.visl.sdu.dk/visl) oferece uma interface gráfica que permite ao usuário analisar, de uma forma interativa – optando entre a análise automática completa e análise manual guiada em vários níveis de complexidade –, exemplos de *corpus* e de materiais de funcionamento livre. A plataforma permite a análise de itens produzidos em diversos idiomas. Embora a descrição gramatical automática seja baseada em CG (*Constraint Grammar*), pode ser transformada em sistemas de anotação especificada para diferentes fins com etiquetas de texto em execução ou códigos de cor no texto. O núcleo de bases de dados de linguagem da VISL são os seus *treebanks*. Os *treebanks* são derivados de pesquisas armazenadas automaticamente.

Essa ferramenta dinamarquesa é utilizada no processo de etiquetagem dos textos do "*Corpus* Linguístico UFSM" por ser ferramenta mais precisa dentre todos os etiquetadores existentes no mundo atualmente. As etiquetas feitas pela VISL são morfossintáticas e são indispensáveis para pesquisadores das mais diversas áreas das Letras. Além disso, é o único etiquetador totalmente gratuito e disponível *on-line* para os pesquisadores do mundo todo. Em contrapartida, a ferramenta analisa apenas um trecho por vez o que dificulta o processo de etiquetagem, pois acaba sendo um processo bastante demorado e cuidadoso.

Para demonstrar como ocorre o processo de etiquetagem na VISL, submeter-se-á à ferramenta um discurso da ex-presidente brasileira Dilma Rousseff conforme será visto na subseção a seguir.

3.2.1 Análise com o auxílio da ferramenta computacional *on-line* VISL

Para exemplificar o uso dessa ferramenta, analisar-se-á a seguir um pequeno excerto extraído do Portal do Planalto no ano de 2014. Trata-se de um trecho do pronunciamento da ex-presidente da República do Brasil, Dilma Rousseff, feito no dia 11 de fevereiro de 2011, em Brasília – DF. Em tal pronunciamento à nação, em cadeia nacional de rádio e TV, a ex-presidente falou sobre a volta às aulas e a Educação no Brasil. Segue abaixo o trecho do discurso político a ser analisado (BRASIL, 2014), e, em seguida, o resultado de sua breve análise com o auxílio da VISL.

> *(1) Nossos jovens estão de volta às aulas. A abertura do ano escolar é sempre uma festa de alegria, de fé e de esperança. É com esse sentimento que saúdo os estudantes, seus pais e, muito especialmente, todos os professores brasileiros.*

Para que sejam compreendidos os resultados da análise feita com a VISL, é necessário que, inicialmente, conheçam-se as abreviaturas utilizadas pelo programa. O Quadro 2 contém a legenda de símbolos e termos utilizados na análise.

Quadro 2 - Legendas de termos utilizados na análise morfossintática com o programa VISL

***2** – dois ou mais
[ ] – traz a forma singular
< > geralmente traz a definição morfológica
1P – 1ª pessoa do plural
1S – 1ª pessoa do singular
3P – 3ª pessoa do plural
3S – 3ª pessoa do singular
ADJ – adjetivo
ADV – advérbio
artd - artigo definido
arti - artigo indefinido
DEM – pronome demonstrativo
DET – determinante
F – feminino
IND – modo verbal indicativo
KC – conectivo
M – masculino

N – **nome,** substantivo
P – plural
PER- período, época, data, estação do ano
POSS – pronome possessivo
PR – tempo verbal presente
PRP – preposição
QUANT – quantidade
S – singular
SUBJ - "subject" = sujeito, substantivo
V – verbo
VFIN - verbo finito
VT - verbo transitivo

Conforme mostra o Quadro 2, para cada classificação morfossintática, a ferramenta VISL apresenta uma legenda, as quais são descritas em ordem alfabética no Quadro 2, como, por exemplo, ADJ para adjetivo e ADV para advérbio. Vale destacar as legendas *2, que significa duas ou mais definições, [ ] (colchetes), que sempre trazem dentro de si a forma singular de cada termo analisado, e, por sua vez, o símbolo < > que vem sempre acompanhado da classe morfológica da palavra analisada.

Como resultado da análise morfossintática *on-line* do trecho em questão, obtiveram-se os resultados ilustrados pelo Quadro 3.

Quadro 3 - Breve análise morfossintática com o auxílio da ferramenta computacional *on-line* VISL

nossos [nosso] <poss 1P> **DET** M P
jovens [jovem] <n> **ADJ** M P
estão [estar] **V** PR 3P IND VFIN
de volta [de=volta] **ADV**
a [a] **PRP**
as [o] <artd> **DET** F P
aulas [aula] **N** F P
.
a [o] <artd> **DET** F S
abertura [abertura] **N** F S
de [de] **PRP**
o [o] <artd> **DET** M S
ano escolar [ano=escolar] <per> **N** M S

é [ser] **V** PR 3S IND VFIN
sempre [sempre] **ADV**
uma [um] <arti> **DET** F S
festa [festa] **N** F S
de [de] **PRP**
alegria [alegria] **N** F S
de [de] **PRP**
fé [fé] **N** F S
e [e] **KC**
de [de] **PRP**
esperança [esperança] **N** F S
é [ser] **ADV**
com [com] **PRP**
esse [esse] <dem> **DET** M S
sentimento [sentimento] **N** M S
que [que] **ADV**
saúdo [saudar] <vt> **V** PR 1S IND VFIN
os [o] <artd> **DET** M P
estudantes [estudante] **N** M P
seus [seu] <poss 3S> **DET** M P
pais [pai] **N** M P
e [e] **KC**
muito [muito] <quant> **ADV**
especialmente [especialmente] **ADV**
todos os [todo=o] <quant> **DET** M P
professores [professor] **N** M P
brasileiros [brasileiro] <*2> **ADJ** M P

De todos os etiquetadores existentes do mundo, a VISL tem se mostrado o mais eficiente com 98% de acertos se comparado com análises manuais. Esses 98% de eficácia podem ser vistos através do Quadro 3. Com exceção de dois itens léxico-gramaticais, o programa mostrou-se certeiro tanto quanto uma análise feita manualmente por um profissional da morfossintaxe. Os dois itens léxico-gramaticais analisados equivocadamente pela ferramenta foram *é* e *que*. Inicialmente, a ferramenta soube identificar a singularidade de ambos, mas acabou por defini-los como ADV (advérbios), quando, na verdade, o *é* classifica-se como verbo *ser* no singular do presente do indicativo e o *que*, neste caso, seria uma conjunção indicando uma oração clivada. A partir dessas constatações, é necessário ressaltar

que, apesar de todas as facilidades da ferramenta, ainda assim faz-se indispensável a conferência das respostas pelo pesquisador de forma que não ocorram erros cruciais nas pesquisas.

4 CONSIDERAÇÕES FINAIS

Com desenvolvimento do presente trabalho foi possível demonstrar como se deu o processo de construção do "*Corpus* Linguístico UFSM", isto é, em que fontes foi feita a coleta do *corpus*, e como aconteceu o seu processo de organização, ou seja, em que formato foram "salvos" os arquivos a fim de facilitar o uso de diversas ferramentas eletrônicas de análise, e como foram dispostos os textos/arquivos.

Pode-se ainda demonstrar, através deste estudo, que (i) a ferramenta de etiquetagem VISL pode facilitar grandemente o trabalho dos pesquisadores, mas (ii) tem apenas 98% de eficácia se comparado a análises manuais, o que justifica a necessidade de revisão dos resultados pelo pesquisador após o término de cada etiquetagem. Além disso, este artigo pretende ressaltar a importância da construção de um vasto acervo de textos para facilitar o desenvolvimento de trabalhos em análises do discurso pelos pesquisadores da área de Letras e por usuários interessados no trabalho com textos, neste caso, das três áreas do discurso (Discurso Acadêmico, Discurso Midiático e Discurso Político).

Atualmente, o "*Corpus* Linguístico UFSM" está construído de mais de dois milhões de palavras, que estão em processo de etiquetagem, e ainda se encontra em expansão. Após a etiquetagem de, no mínimo, dois milhões de palavras, o *corpus* será disponibilizado em um site institucional.

5 REFERÊNCIAS

BERBER SARDINHA, T. **Linguística de *Corpus***. Barueri: Editora Manole, 2004.

BICK, E. **Visual Interactive Sintax Learning**. Denmark: University of Southern Denmark, Institute of Language and Communication, 1996. Disponível em: <http://visl.sdu.dk/visl/>. Acesso em: jul. 2017.

BIBER, D.; CONRAD, S.; REPPEN, R. ***Corpus* linguistics**: Investigating language structure and use. Cambridge: Cambridge University Press, 1998.

BLOOR, T., BLOOR, M. **The functional analysis of English** – a hallidayan approach. London: Edward Arnold, 1995.

BRASIL. **Portal do Planalto**. Pronunciamento à nação da Presidenta da República, Dilma Rousseff, em cadeia nacional de rádio e TV em 10 de fevereiro de 2011. Brasília, 2014. Disponível em: < http://www2.planalto.gov.br> Acesso em: nov. 2016.

FUZER, C.; CABRAL, S. R. S. **Introdução à Gramática Sistêmico-Funcional em Língua Portuguesa.** Campinas: Mercado de Letras, 2014.

HALLIDAY, M. A. K. **An introduction to functional grammar**. 2nd ed. London: Edward Arnold, 1994.

___. **The Language of science.** First published. London: Continuum, 2004.

HALLIDAY, M. A. K.; MATTHIESSEN, C. M. I. M. **Construing experience through meaning:** a language-based approach to cognition. London and New York: Continuum, 1999.

___. **An introduction to functional grammar**. 3rd ed. London: Edward Arnold, 2004.

___. **An introduction to functional grammar**. 4th ed. London: Hodder Education, 2014.

HALLIDAY, M. A. K. Context of situation. In: HALLIDAY, M. A. K.; HASAN, R. **Language, context and text:** aspects of a language in a social-semiotic perspective. Oxford: Oxford University, 1985.

LAGE, N. **Linguagem jornalística.** São Paulo: Ática, 1993.

MARTIN, J. R.; MATTHIESSEN, C. M. I. M.; PAINTER, C. **Working with functional grammar.** London: Edward Arnold, 1997.

MILLER, C. R. **Gênero textual, agência e tecnologia**. São Paulo: Parábola Editorial, 2012.

RABAÇA, C. A.; BARBOSA, G. **Dicionário de comunicação**. 2. ed. rev. Rio de Janeiro: Elsevier, 2001.

SCOTT, M. **Programa WordSmith Tools** (version 7.0) [computer software]. Oxford: Oxford University Press, 2016.

SPITULNIK, D. Anthropology and mass media. **Annual Review of Anthropology**, v. 22, p. 293-315, 1993. Disponível em: http://www.annualreviews.org/doi/abs/10.1146/annurev.an.22.100193.001453?prevSearch=authors%2528spitulnik%2529&searchHistoryKey=. Acesso em: set. 2015.

THOMPSON, G. **Introducing Functional Grammar**. 2nd ed. London: Arnold, 2004.

TOGNINI-BONELLI, E. ***Corpus* linguistics at work**. Amsterdã/Atlanta: John Benjamins, 2001.