Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Caracterização linguística de sumários humanos multidocumento: explorando o nível lexical

Vanessa Marcasso
Ariani Di Felippo

NILC-TR-14-02

Setembro, 2014

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste relatório, descreve-se uma investigação sobre características lexicais de textos-fonte e sumários humanos multidocumento. Com o objetivo de estender o conhecimento atual sobre as características lexicais dos referidos documentos (textos-fonte e sumários), aplicou-se uma série de medidas estatísticas ao CSTNews, *corpus* multidocumento jornalístico em português. As medidas estatísticas (i) razão *type-token*, (ii) número de palavras (tamanho), (iii) incidência de palavras de conteúdo, (iv) frequências e (v) índice Flesch foram aplicadas ao CSTNews por meio do ambiente *Coh-Metrix-Por*, disponível no portal do projeto PorSimples. A pesquisa ora descrita foi realizada em uma iniciação científica que compreendeu o período de 01/09/2013 a 31/08/2014.

Este trabalho conta com o apoio
financeiro da FAPESP (2013/12524-4).

1. Introdução

As pesquisas sobre as características dos sumários humanos multidocumento (ou seja, produzidos a partir de uma coleção de textos que abordam um mesmo assunto) começaram a ser realizadas recentemente, motivadas, sobretudo, pelo interesse na produção automática de sumários linguisticamente motivados.

Nesse contexto, destacam-se as pesquisas sobre os sumários jornalísticos do tipo informativo e genérico, os quais veiculam, de forma concisa e coerente/coesa, o conteúdo principal de uma coleção de notícias jornalísticas (sobre um mesmo fato) de tal forma que uma audiência genérica pode dispensar a leitura das mesmas (MANI, 2001).

Tais pesquisas evidenciaram que os sumários são comumente compostos por (NENKOVA, 2006; CAMARGO, 2013, RASSI et al., 2013): (i) informações que se localizam na parte inicial dos textos-fonte; (ii) a informação mais redundante da coleção, pois esta é tida como a mais relevante, (iii) informações provenientes de um dos textos-fonte da coleção em específico, entre outras.

O trabalho de Camargo (2013) destaca-se por ser o primeiro a investigar diversas propriedades de um *corpus* em português. É por causa dele que hoje se sabe mais sobre os sumários do único *corpus* multidocumento de referência do português, o CSTNews (CARDOSO et al., 2011). Sobre o nível lexical, Camargo revelou que as palavras mais frequentes da coleção estão presentes em seu respectivo sumário.

Inserido no cenário do projeto SUSTENTO (FAPESP 2012/13246-5/ CNPq 483231/2012-6), cujo objetivo é o de gerar subsídios linguísticos para a Sumarização Automática Multidocumento (SAM) em português, este trabalho deu continuidade ao de Camargo (2013), através da descrição e análise das características dos sumários multidocumento e seus respectivos textos-fonte, em nível lexical.

Na Seção 2, discorre-se sobre conceitos básicos de SAM e revisam-se alguns trabalhos em que as propriedades lexicais de sumários multidocumento são enfocadas. Na Seção 3, apresenta-se o CSTNews, *corpus* multidocumento cujos textos-fonte e sumários foram aqui investigados. Na Seção 4, descreve-se a aplicação de várias medidas estatísticas lexicais ao *corpus* CSTNews. Na Seção 5, apresenta-se uma interpretação para os resultados das medidas em função das agências jornalísticas dos textos-fonte. Na Seção 6, são apresentadas as considerações finais deste trabalho e os apontamentos para trabalhos futuros.

2. Revisão da literatura

Os assuntos abordados na revisão da literatura foram: (i) as propriedades dos sumários multidocumento e de seus textos-fonte e (ii) medidas estatísticas que capturam a riqueza lexical e a complexidade ou inteligibilidade de textos.

2.1. Os sumários multidocumento e as suas características

Os textos jornalísticos são da "ordem do relatar" e do "domínio social da memorização e documentação das experiências humanas" (DOLZ, SCHNEWLY, 2004). Nesse grupo, citam-se as notícias e os seus sumários informativos e genéricos, que veiculam o conteúdo principal dos textos-fonte a uma audiência genérica.

As características linguísticas desses sumários, quando produzidos manualmente a partir de várias notícias sobre um mesmo assunto (multidocumento), foram alvo de algumas pesquisas sistemáticas baseadas em *corpus*, motivadas principalmente pelo interesse na produção automática de sumários com as mesmas propriedades dos elaborados por humanos (NENKOVA, 2006; CAMARGO, 2013; RASSI et al., 2013).

Um dos aspectos dos sumários multidocumento é a presença de informação que ocorre em certas posições dos textos-fonte. Os sumários jornalísticos marcam-se pela ocorrência de informações expressas no início do texto-fonte (CAMARGO, 2013).

Isso se deve ao fato de que as notícias são compostas por: (i) título, (ii) *lead*, que corresponde ao primeiro parágrafo do texto, e (iii) corpo do texto, que abrange os demais parágrafos, os quais desenvolvem os elementos informativos referidos no *lead*.

O *lead* é a informação principal, a qual é expressa com o intuito de instigar o leitor. Logo no início, o leitor tem o essencial da informação, que engloba **quem** (sujeito), **o que** (fato/acontecimento), **quando** (tempo), **por quê** (causa/motivo/finalidade), **como** (modo/maneira) e **onde** (lugar).

Essa estrutura típica é chamada de "pirâmide invertida" (LAGE, 2002), ilustrada na Figura 1. Por expressar o conteúdo principal de um texto jornalístico, a informação localizada no início frequentemente compõe o sumário.

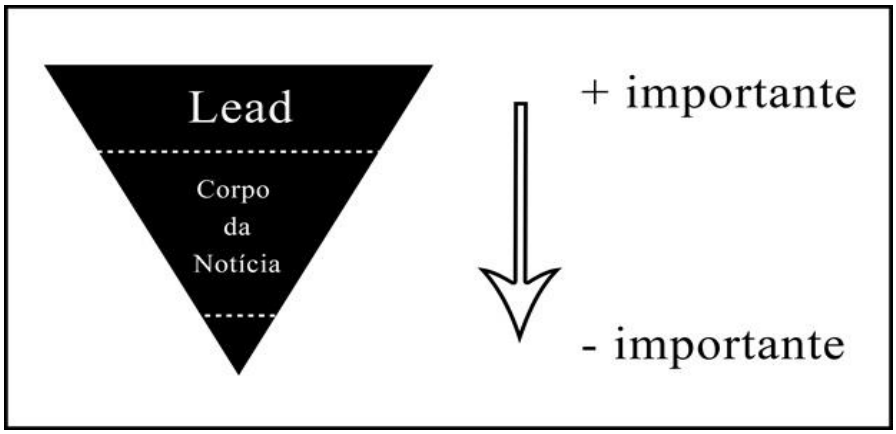

Figura 1: Modelo estrutural da pirâmide invertida.
Fonte: Elaborada com base em Lage (2002)

Outra característica dos sumários multidocumento é a veiculação da informação mais redundante da coleção, ou seja, a informação mais repetida nos textos-fonte. Essa caraterística foi comprovada por Nenkova (2006) e Camargo (2013).

Nenkova (2006) constatou essa característica ao verificar que as sentenças que compõem esse tipo de sumário apresentam as palavras (de conteúdo) mais frequentes da coleção. Utilizando um *corpus* em inglês de 30 coleções, cada uma composta por 10 textos jornalísticos compilados do jornal *The New York Times*, 10 sujeitos produziram um sumário multidocumento para cada coleção com aproximadamente 100 palavras. Desse experimento, a autora comprovou que em média 94,66% das palavras mais frequentes de uma coleção estão presentes nos respectivos sumários.

Camargo (2013) constatou a presença da informação mais redundante de forma diferente e em um *corpus* em língua portuguesa. No caso, a autora utilizou o CSTNews (CARDOSO *et al*., 2011), *corpus* multidocumento com 50 coleções, cada uma delas com em média 3 notícias e um sumário manual multidocumento.

As sentenças dos sumários foram manualmente indexadas às sentenças de origem dos textos-fonte. Após essa indexação, várias propriedades das sentenças dos textos-fonte alinhadas[1] aos sumários foram descritas. No Quadro 2, observa-se a caracterização das sentenças S1 e S2 do documento D1 da coleção C1 do CSTNews. Os atributos, que se dividem em superficiais, estrutural, profundo e extralinguístico, sintetizam propriedades potencialmente relevantes para a caracterização dos sumários.

Quadro 2: Exemplificação da caracterização das sentenças.

Alinhamento		**Atributos**								
		Superficiais			Estrutural	Profundos (no. de relações CST)				Extralinguístico
Sentença (texto-fonte)	Sumário	**Tamanho** da sentença	**Frequência** das palavras de conteúdo	**Palavra-chave** (10 mais frequentes da coleção)	**Localização**	**Forma**	**Redundância**	**Complemento**	**Contradição**	Fonte
S1_D1_C1	sim	0.47	0.63	0.71	começo	1	0,66	0.57	0	Folha de S. Paulo
S2_D1_C1	não	0.68	0.6	0.57	meio	1	0.66	1	0	Folha de S. Paulo

Os atributos profundos, em especial, capturam o número de relações CST que as sentenças alinhadas (e não-alinhadas) possuem, pois, em cada coleção do CSTNews, as sentenças dos textos-fonte estão conectadas por relações semântico-discursivas da teoria CST (*Cross-document Structure Theory*) (RADEV, 2000), como ilustrado na Figura 2, elaborada com base em Maziero (2012).

[1] As sentenças não-alinhadas, ou seja, cujo conteúdo não foi selecionado para compor os sumários, também foram caracterizadas por motivos de comparação.

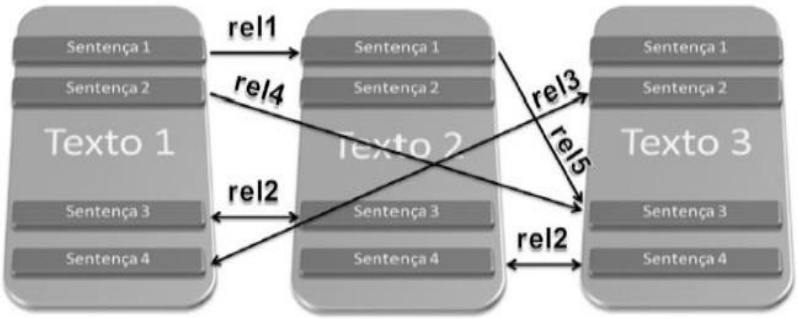

Figura 2: Esquema genérico de alinhamento multidocumento.
Fonte: Maziero (2012).

A CST prevê um conjunto de relações que capturam redundância, complementaridade, contradição e variação de escrita entre pares de sentenças provenientes de textos distintos sobre mesmo assunto. No CSTNews, as sentenças encontradas foram: (i) *identity, equivalence, summary*, *subsumption* e *overlap* (redundância), (ii) *historial background, follow-up* e *elaboration* (complemento), (iii) *contradiction* (contradição), (iv) *citation, attribution, madality, indirect speech* e *translation* (variação de forma).

Com base na caracterização das sentenças alinhadas, Camargo não só caracterizou indiretamente as sentenças dos sumários (p.ex.: quanto ao seu tamanho médio, frequência média das palavras de conteúdo, localização no texto-fonte, etc.) como também identificou as propriedades de destaque.

Dentre elas, Camargo verificou que todas as sentenças dos textos-fonte que possuem 5 relações CST de redundância foram alinhadas aos sumários, o que comprova que a informação mais redundante da coleção compõe um sumário multidocumento correspondente.

Ademais, Camargo (2013) verificou que as informações constitutivas dos sumários são provenientes de um texto-fonte de preferência, dentre os vários da coleção. No caso, 56% das sentenças alinhadas aos sumários do CSTNews são dos textos da *Folha de São Paulo*. A razão para isso é difícil de ser apontada. Os humanos podem ter selecionado essa fonte por causa, por exemplo, do seu prestígio.

Outra característica comprovada por meio de estudos sistemáticos é a de que os sumários multidocumento apresentam conjuntos específicos de "aspectos" (isto é, unidades básicas de informação ou discursivas) em função de sua categoria/domínio (p.ex.: OWCZARZAK, DANG, 2011; e LI *et al*., 2011, RASSI et al, 2013). Os aspectos textuais, como componentes discursivos, remontam a Swales (1990) que, em uma análise do gênero artigo acadêmico, identificou a estrutura retórica ou esquemática dividida nos componentes: introdução, métodos, resultados e discussão.

Para o português, o estudo pioneiro desenvolvido por Rassi et al (2013) utilizou o *corpus* CSTNews (CARDOSO et al., 2011), cujas 50 coleções são das categorias "mundo" (14), "cotidiano" (14), "política" (10), "esporte" (10), "dinheiro" (1) e "ciência" (1).

Para tanto, os sumários foram manualmente anotados com 16 aspectos, representados por etiquetas em inglês, como *who_agent* (agentes), *who_affected* (pacientes), *what* (objetos genéricos/eventos), *when* (datas), etc. Da anotação, Rassi et al. identificaram organizações prototípicas dos aspectos nos sumários de cada categoria. Por exemplo, a maioria dos sumários de "esporte" veiculam os aspectos *who_agent, what, score, consequence, situation, comment, when* e *where*, enquanto a maioria dos sumários da categoria "política" veicula os aspectos *who_agent, what, when* e *why*.

Das pesquisas sobre as características dos sumários multidocumento, destacam-se as de Camargo (2013) e Rassi et al. (2013), pois consistem em investigações sistemáticas de *corpus* multidocumento em português. É por causa delas que hoje se sabe um pouco mais sobre o tipo de sumário em questão.

Quanto às características do nível lexical, o trabalho de Camargo (2013) demonstrou, por exemplo, que aproximadamente 70% das sentenças dos textos-fonte constituídas por palavras de conteúdo bastante frequentes ou pelas mais frequentes da coleção (ou seja, pelas palavras-chave) foram alinhadas aos sumários. Isso significa que um sumário multidocumento do CSTNews é composto por sentenças constituídas pelas palavras mais frequentes de sua respectiva coleção.

Buscando refinar o conhecimento sobre as características lexicais dos sumários, descrevem-se, a seguir, estatísticas lexicais capazes de revelar a riqueza lexical e o grau de complexidade ou inteligibilidade dos textos.

2.2. Medidas estatísticas lexicais

A medida estatística *Type-token ratio* (SCOTT, 2014; MANNING, SCHÜTZE, 1999; BERBER SARDINHA, 2004) e o pacote de medidas Coh-Metrix-Port destacam-se no âmbito da Linguística de Corpus[2] por capturar a diversidade/riqueza vocabular de um texto e a complexidade/inteligibilidade textual, respectivamente.

a) *Type-token ratio*

Type-token ratio (razão TT) determina a diversidade vocabular de um texto. Partindo-se do princípio de que um texto é constituído por palavras cuja maioria é utilizada várias vezes e apenas uma parte é diferente, a razão TT calcula a porcentagem de tipos de palavras face ao total de palavras do texto (MANNING, SCHÜTZE, 1999).

Assim sendo, *token* equivale a uma palavra corrida (do inglês, *running words*) e, por conseguinte, cada ocorrência de uma palavra em um texto conta como um *token* (item). *Type* (forma) equivale a uma palavra distinta. Para ilustrar, considera-se a sentença (1) "*A Joana conversou com a Maria. Já a Patrícia, conversou com a Joana, mas não falou com a Maria*". Nela, há 19 *tokens* e 10 *types*, como descrito no Quadro 3.

[2] Área de pesquisa que se ocupa da "coleta e exploração de *corpora*, ou conjunto de dados linguístico-textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística" (BERBER SARDINHA, 2004, p.3).

A partir da identificação dos *tokens* e *types* de um texto, calcula-se a razão TT. Para tanto, divide-se o total de *types* pelo total de *tokens* (BERBER SARDINHA, 2004). No caso da sentença em (1), a razão TT é de aproximadamente 0,52 (10/19 = 0,526). A razão TT também pode ser expressa em porcentagem, dividindo-se o total de *types* pelo resultado da divisão de *tokens* por cem (*type* / (*token* / 100)). No caso de (1), tem-se a razão TT de aproximadamente 52,6% (10 / (19 / 100) = 52,6).

Quadro 3: Exemplificação dos conceitos *type* e *token.*

Token	***Type***
a	a
a	
a	
a	
a	
com	com
com	
com	
conversou	conversou
conversou	
falou	falou
já	já
joana	joana
joana	
maria	maria
maria	
mas	mas
não	não
patrícia	patrícia
Total: 19	**Total:** 10

Como mencionado, a razão TT é uma medida que determina a diversidade vocabular de um texto. No caso, quanto mais alto é o valor da razão, mais palavras diferentes o texto contém, ou seja, mais diversificado é o seu vocabulário. Por outro lado, quanto mais baixa é a razão TT, menor diversidade lexical o texto apresenta, posto que o número de repetições é elevado. Assim, considera-se que o texto é menos rico no nível lexical (BERBER SARDINHA, 2004).

Há também a *standardised type-token ratio* (razão *type-token* padronizada ou razão STT), que é calculada em intervalos regulares (SCOTT, 2014). O resultado é a média de todas as densidades parciais, calculadas tantas quantas forem as vezes em que o intervalo esteja presente ao longo de um texto ou de um conjunto de textos.

Por exemplo, ao se dividir o texto em (1) na metade, tem-se "*A Joana conversou com a Maria. Já a Patrícia,"* (Trecho 1) e "*conversou com a Joana, mas não falou com a Maria"* (Trecho 2), cujos *types* e *tokens* estão no Quadro 3. O

Trecho 1 consiste de 7 *types* e 9 *tokens*, com uma razão TT de 77,7 (7 / (9 / 100) = 77,7), e o Trecho 2 possui 8 *types* e 10 *tokens*, com uma razão TT de 80,0 (8 / (10 / 100) = 80) (Quadro 4). Tirando-se a média aritmética, chega-se a 78,8 (77,7 + 80 / 2), que é o valor da razão STT.

Quadro 4: *Types* e *tokens* dos Trechos 1 e 2.

Trecho 1		**Trecho 2**	
Type	***Token***	***Type***	***Token***
a	a	a	a
com	a	com	a
conversou	a	conversou	com
já	com	falou	com
joana	conversou	joana	conversou
maria	já	maria	falou
patrícia	joana	mas	joana
	maria	não	maria
	patrícia		mas
			não
Total: 7	**Total:** 9	**Total**: 8	**Total:** 10

Segundo Berber Sardinha (2004), a forma padronizada permite a comparação de textos a partir da determinação de um intervalo comum para o cálculo, pois textos maiores, por sua natureza, apresentam mais repetições de palavras (ou um número maior de *tokens*) e, por isso, tendem a possuir valores para essa razão mais baixos do que textos curtos. No caso do texto em (1), por exemplo, a razão TT foi de 52,6 enquanto que a razão STT foi de 78,8.

Em outras palavras, a razão STT é empregada para neutralizar a influência do tamanho do texto na comparação da razão TT, pois a TT é sensível ao tamanho e, por isso, não é confiável para comparações entre textos de tamanhos diferentes.

As razões TT e STT podem ser obtidas de forma automática por meio do uso de *softwares* como o *WordSmith Tools*[3] (WS) e o *Simple Concordance Program*[4] (SCP).

b) *O Coh-Metrix e o Coh-Metrix Port*

O *Coh-Metrix* (do inglês, *cohesion metrics*) (CM), é uma das várias tecnologias do Processamento Automático das Línguas Naturais (PLN)[5] disponíveis para a descrição e análise textual. O CM foi elaborado por pesquisadores da Universidade de Memphis (EUA), sendo seu propósito calcular índices de coesão/coerência textual

[3] http://www.lexically.net/wordsmith/
[4] http://www.textworld.com/scp
[5] O PLN é uma área de pesquisa em que se busca desenvolver sistemas computacionais capazes de realizar tarefas linguísticas específicas como tradução, sumarização, etc. (DIAS-DA-SILVA et al., 2007).

por meio de várias medidas estatísticas lexicais, sintáticas, semânticas e referenciais, as quais indicam a complexidade ou inteligibilidade dos textos. Mais de 500 métricas estão disponíveis na versão restrita do *Coh-Metrix* e 60 estão disponíveis na versão gratuita[6].

A partir do *Coh-Metrix* em inglês, uma iniciativa de adaptação para o português brasileiro das sessenta métricas gratuitas surgiu no Projeto PorSimples, cujo objetivo era o de identificar índices de complexidade textual para simplificação de textos e facilitação do acesso à informação a analfabetos funcionais e pessoas com deficiências cognitivas. Tal iniciativa resultou na versão em português do *Coh-Metrix* original, que foi denominada Coh-Metrix-Port (CMP) e está disponível no site do PorSimples.

O Coh-Metrix-Port dispõe hoje de 35 métricas adaptadas para o português, das sessenta métricas originais do inglês. As 35 métricas abrem, como bem salienta Pasqualini (2011), um amplo universo de possibilidades para os estudos linguísticos no que tange à complexidade ou inteligibilidade textual. As métricas do Coh-Metrix-Port podem ser classificadas quanto ao nível linguístico em que atuam, a saber: lexical, sintático e semântico[7]. As principais medidas lexicais estão descritas no Quadro 5.

Quadro 5: Métricas lexicais do Coh-Metrix-Port.

Nível	**Métricas**
Lexical	Número de palavras
	Incidência de palavras de conteúdo
	Frequências
	Índice *Flesch*

A métrica "número de palavras" captura a quantidade de palavras contidas em um texto ou *corpus* e equivale ao número de *tokens.* Por exemplo, na sentença "*Acessório utilizado por adolescentes, o boné é um dos itens que compõem a vestimenta idealizada pela proposta*", tem-se 17 palavras (*tokens*).

A "incidência de palavras de conteúdo" captura a quantidade desse tipo de palavra no texto/*corpus* em função do número total de palavras (*tokens*) do texto/*corpus.* No exemplo, na sentença "***Acessório utilizado*** *por* ***adolescentes****, o* ***boné é*** *um dos* ***itens*** *que* ***compõem*** *a* ***vestimenta idealizada*** *pela* ***proposta***", há 10 palavras de conteúdo e 17 palavras no total, o que gera a "incidência de palavras de conteúdo" de 588,23 (10 / (17 /1000)). Essa métrica pode ser especificada para cada classe de palavra de conteúdo. Assim, o Coh-Metrix-Port pode calcular também as métricas "incidência de substantivo", "incidência de adjetivos", "incidência de verbos" e "incidência de advérbios". Na sentença "***Acessório*** *utilizado por* ***adolescentes****, o* ***boné*** *é um dos* ***itens*** *que compõem a* ***vestimenta*** *idealizada pela*

[6] http://cohmetrix.memphis.edu/cohmetrixpr/index.html
[7] As métricas que atuam no nível referencial ainda não foram implementadas no Coh-Metrix-Port.

proposta", por exemplo, as palavras em negrito totalizam 6 substantivos, que gera a "incidência de substantivo" de 352,94 (6 / (17 /1000)).

A métrica "frequências"[8] expressa a média de todas as frequências das palavras de conteúdo do texto ou *corpus*. O valor da frequência das palavras é retirado da lista de frequências do *corpus* Banco de Português[9] (BP). Por exemplo, de acordo com a lista frequências do BP, o valor médio das frequências das palavras de conteúdo em negrito da sentença "***Acessório utilizado*** *por* ***adolescentes****, o* ***boné é*** *um dos* ***itens*** *que* ***compõem*** *a* ***vestimenta idealizada*** *pela* ***proposta***" é 32432,7.

Quanto ao índice *Flesch* (CMP), ressalta-se que este foi proposto por Rudolph Flesch, que, tendo fugido da Europa nazista durante a guerra e se repatriado nos Estados Unidos, defendia o uso de um inglês simplificado (*plain English*), ou seja, claro, objetivo e sem ambiguidades em documentos oficiais para fácil compreensão de todos os cidadãos e em determinadas situações de ensino. Para tanto, Flesch propôs um índice de determinação da inteligibilidade que ficou conhecido como "índice *Flesch*". A fórmula constante no Coh-Metrix-Port para o Índice de Legibilidade *Flesch* (ILF) é a descrita em (2), sendo ASL o número de palavras dividido pelo número de sentenças e ASW o número de sílabas dividido pelo número de palavras. Os resultados podem ser agrupados em 4 categorias de complexidade (SCARTON et al., 2010) (Quadro 6).

(2) $$\mathbf{ILF = 248.835 - (1.015 \times ASL) - (84.6 \times ASW)}$$

Quadro 6: Categorias de complexidade textual baseadas no índice *Flesch.*

Categoria	ILF	Descrição
Muito fácil	75-100	Textos adequados para leitores com nível de escolaridade até a quarta série do ensino fundamental
Fácil	50-75	Textos adequados a alunos com escolaridade até a oitava série do ensino fundamental
Difícil	25-50	Textos adequados para alunos cursando o ensino médio ou universitário
Muito difícil	0-25	Textos adequados apenas para áreas acadêmicas específicas

3. Seleção e pré-processamento do *corpus*

O CSTNews (CARDOSO et al., 2011) é um *corpus* multidocumento composto por 50 coleções de textos jornalísticos, sendo que cada coleção versa sobre um mesmo tópico. Os textos são do gênero discursivo "notícias jornalísticas", cujas principais propriedades são (i) documentar as experiências humanas vividas (domínio social) e

[8] A "frequência" não expressa o número de ocorrências de cada palavra no texto/*corpus*, mas a frequência dessas palavras em um *corpus* de referência, o *corpus* Banco do Português.

[9] O Coh-Metrix-Port retira as frequências do Banco de Português (BP). Em 2004, o BC possuía 223 milhões de palavras (*tokens*). Mais informações podem ser encontradas no endereço: http://www2.lael.pucsp.br/corpora/bp/index.htm

(ii) representar pelo discurso as experiências vividas, situadas no tempo (capacidade da linguagem) (BARBOSA, 2001; DOLZ; SCHNEWLY, 2004; LAGE, 2002).

Cada coleção contém basicamente: (i) 2 ou 3 textos sobre um mesmo assunto, compilados de diferentes fontes jornalísticas e com tamanhos similares, (ii) sumários humanos mono e multidocumento, (iii) sumários automáticos multidocumento, e (v) diversas anotações, como a que se baseia na CST, dentre outros dados.

As fontes jornalísticas das quais os textos foram compilados correspondem aos principais jornais online do Brasil: *Folha de São Paulo* (Folha), *Estado de São Paulo* (Estadão), *Jornal do Brasil* (JB), *O Globo* e *Gazeta do Povo* (GP). A coleta manual foi entre agosto a setembro de 2007. As coleções possuem em média 42 sentenças (de 10 a 89) e os sumários humanos multidocumento possuem em média 7 sentenças (de 3 a 14).

Ademais, as coleções estão categorizadas pelos rótulos das "seções" dos jornais dos quais os textos foram compilados. Assim, o *corpus* é composto por coleções das categorias: "esporte" (10 coleções), "mundo" (14 coleções), "dinheiro" (1 coleção), "política" (10 coleções), "ciência" (1 coleção) e "cotidiano" (14 coleções).

Os sumários humanos multidocumento foram construídos de forma abstrativa, ou seja, com reescrita do conteúdo dos textos-fonte. Além disso, a produção dos mesmos foi guiada por uma taxa de compressão de 70%. Consequentemente, os sumários contêm, no máximo, 30% do número de palavras do maior texto-fonte da coleção. Do ponto de vista da audiência, os sumários do CSTNews são genéricos, e, do ponto de vista funcional, são informativos, pois contemplam as informações principais de seus textos-fonte, substituindo a leitura dos mesmos.

A seguir, apresenta-se a descrição e análise lexical dos sumários e textos-fonte do CSTNews, que consistiu no cálculo da razão TT e da complexidade/inteligibilidade textual. Para tanto, nenhum pré-processamento do *corpus* foi necessário.

4. Descrição/análise lexical dos sumários/textos-fonte

4.1. A razão TT

Tendo em vista que os textos-fonte de cada coleção do CSTNews possuem tamanhos similares (mas não iguais), optou-se pelo cálculo da razão TT e não da STT. Ademais, optou-se pelo uso dos *softwares* livres *Simple Concordance Program* (SCP) e Coh-Metrix-Port (CMP).

O SCP fornece vários recursos, por meio de uma interface *off-line* amigável (Figura 3).

Figura 3: Interface do *Simple Concordance Program.*

Nela, vê-se que há 3 abas: *Concordance*, *Word List* e *Statistics*. Na aba *Concordance,* a ferramenta gera uma listagem de contextos (palavras ao redor) nos quais um dado item (p.ex.: uma palavra isolada) ocorre no texto ou *corpus*. Na aba *Word List,* o SCP exibe a listagem de palavras do texto/*corpus* juntamente com suas frequências no texto/*corpus*. Na aba *Statistics,* em especial, gera o número de *types* e *tokens* e a razão TT.

Os textos-fonte e o sumário multidocumento de cada uma das 50 coleções do CSTNews foram submetidos individualmente ao SCP e os dados gerados na aba *Statistics* foram organizados em planilhas no formato *xls* (*Excel*). Para cada coleção do CSTNews, elaborou-se uma planilha como a ilustrada na Tabela 1, em que os *types* e *tokens* de cada texto-fonte e sumário foram discriminados. No caso, a Tabela 1 descreve apenas parte da tabela referente à coleção C1 do CSTNews.

Tabela 1: Recorte da planilha com os dados gerados SCP para a coleção C1 do CSTNews.

D1_C1_Folha		**D2_C1_Estadão**		**D3_C1_JB**		**Sumário**	
Type	**Token**	**Type**	**Token**	**Type**	**Token**	**Type**	**Token**
120	**180**	**91**	**127**	**89**	**125**	**50**	**60**
de	10	de	7	de	7	avião	3
a	8	o	5	o	5	de	3
uma	6	da	4	da	4	e	3
congo	5	e	4	e	4	a	2
em	5	avião	3	avião	3	após	2
do	4	do	3	do	3	o	2

Na sequência, o processo de obtenção da razão TT foi repetido utilizando-se, agora, o CMP. Na Tabela 2, tem-se os valores da razão TT para cada texto-fonte e sumário do CSTNews obtidos por meio do SCP e CMP.

Tabela 2: Razão TT dos textos-fonte e sumários do CSTNews obtidas pelo SCP e CMP.

Cluster	**Texto**	**TT SCP**	**TT CMP**
C1	D1_C1_Folha	0,666666667	0,855769
	D2_C2_Estadao	0,716535433	0,87013
	D3_C1_JB	0,71	0,866667
	Sumário	0,833333333	0,941176
C2	D1_C2_Folha	0,472154964	0,603774

	D2_C2_Estadao	0,422705314	0,518018
	D4_C2_Globo	0,682352941	0,846939
	Sumário	0,611510791	0,763889
C3	D1_C3_Folha	0,490968801	0,730435
	D2_C3_Estadao	0,596153846	0,72093
	D3_C3_JB	0,538283063	0,77381
	Sumário	0,612244898	0,823009
C4	D1_C4_Folha	0,558490566	0,700599
	D2_C4_Estadao	0,623318386	0,790698
	D3_C4_OGlobo	0,589285714	0,835227
	Sumário	0,717948718	0,95
C5	D1_C5_JB	0,650684932	0,849462
	D2_C5_GPovo	0,578199052	0,792157
	Sumário	0,742424242	0,910256
C6	D1_C6_Folha	0,489552239	0,683938
	D2_C6_Estadao	0,534482759	0,721393
	D3_C6_JB	0,569565217	0,736434
	Sumário	0,643478261	0,791045
C7	D1_C7_Estadao	0,547368421	0,772727
	D2_C7_OGlobo	0,671568627	0,873874
	Sumário	0,709677419	0,906667
C8	D1_C8_Estadao	0,69047619	0,895833
	D2_C8_Oglobo	0,704142012	0,872093
	D3_C8_JB	0,591269841	0,845528
	Sumário	0,729411765	0,886792
C9	D1_C9_Folha	0,562248996	0,722222
	D2_C9_Estadao	0,535962877	0,781377
	D3_C9_Oglobo	0,632142857	0,812865
	Sumário	0,748251748	0,920455
C10	D3_C10_OGlobo	0,72826087	0,843137
	D4_C10_JB	0,553648069	0,762452
	D5_C10_Gpovo	0,547029703	0,797357
	Sumário	0,666666667	0,806818
C11	D3_C11_Oglobo	0,578740157	0,771242
	D4_C11_JB	0,507462687	0,724138
	D5_C11_GPovo	0,675257732	0,838983
	Sumário	0,652694611	0,813084
C12	D1_C12_Folha	0,576323988	0,764368
	D12_C12_Estadao	0,565625	0,764368
	D3_C12_GPovo	0,601208459	0,831579
	Sumário	0,75	0,870968
C13	D1_C13_Folha	0,567213115	0,812865
	D2_C13_Estadao	0,568561873	0,821429
	D3_C13_Gpovo	0,561797753	0,760766
	Sumário	0,783333333	0,957747
C14	D2_C14_Estadao	0,631782946	0,865248
	D3_C14_Oglobo	0,564912281	0,765101

	D4_C14_JB	0,677083333	0,872549
	Sumário	0,721649485	0,854167
C15	D2_C15_Estadao	0,695652174	0,888889
	D3_C15_Oglobo	0,600760456	0,809211
	D4_C15_JB	0,689189189	0,872093
	Sumário	0,72826087	0,92
C16	D1_C16_Folha	0,548543689	0,768293
	D2_C16_Estadao	0,579158317	0,809211
	D3_C16_Oglobo	0,820512821	0,929577
	Sumário	0,68627451	0,942529
C17	D1_C17_Folha	0,507718696	0,768769
	D2_C17_Estadao	0,571052632	0,770563
	Sumário	0,690217391	0,883495
C18	D1_C18_Folha	0,539170507	0,772358
	D2_C18_Estadao	0,489942529	0,730479
	D3_C18_GPovo	0,69047619	0,85
	Sumário	0,632768362	0,757009
C19	D1_C19_Folha	0,701492537	0,8
	D2_C19_Estadao	0,695652174	0,88
	Sumário	0,779661017	0,916667
C20	D1_C20_Folha	0,502212389	0,73444
	D4_C20_JB	0,714285714	0,862069
	D5_C20_Gpovo	0,589164786	0,825911
	Sumário	0,656934307	0,861111
C21	D1_C21_Estadao	0,584415584	0,73494
	D2_C21_Oglobo	0,479338843	0,682482
	D3_C21_JB	0,53539823	0,706349
	Sumário	0,620915033	0,790698
C22	D2_C22_Estadao	0,511688312	0,702703
	D4_C22_JB	0,574257426	0,738372
	D5_C22_Gpovo	0,582089552	0,837037
	Sumário	0,71942446	0,855422
C23	D1_C23_Folha	0,662650602	0,805825
	D2_C23_Estadao	0,562189055	0,786957
	Sumário	0,729508197	0,929577
C24	D2_C24_Estadao	0,546428571	0,7125
	D3_C24_JB	0,780487805	0,958333
	D4_C24_Gpovo	0,625615764	0,839286
	Sumário	0,714285714	0,877551
C25	D1_C25_Folha	0,498214286	0,728435
	D4_C25_JB	0,556935818	0,827451
	D5_C25_Gpovo	0,527237354	0,775168
	Sumário	0,635220126	0,775281
C26	D1_C26_Folha	0,508250825	0,716667
	D4_C26_JB	0,585253456	0,804167
	D5_C26_GPovo	0,552197802	0,75576
	Sumário	0,664835165	0,81982

C27	D1_C27_Folha	0,500963391	0,738255
	D2_C27_Estadao	0,56377551	0,775862
	D4_C27_JB	0,432907348	0,622951
	Sumário	0,563451777	0,736842
C28	D1_C28_Folha	0,598290598	0,80303
	D2_C28_Estadao	0,551236749	0,798611
	D3_C28_OGlobo	0,582474227	0,795918
	Sumário	0,78313253	0,955556
C29	D1_C29_Folha	0,580121704	0,777027
	D2_C29_OGlobo	0,59347181	0,79703
	D3_C29_GPovo	0,59347181	0,79703
	Sumário	0,705882353	0,862069
C30	D1_C30_Folha	0,454545455	0,619632
	D2_C30_Estadao	0,5	0,661202
	D3_C30_OGlobo	0,464373464	0,657005
	Sumário	0,690140845	0,907895
C31	D1_C31_Folha	0,777777778	0,93617
	D2_C31_Estadao	0,694029851	0,881579
	Sumário	0,74	0,9
C32	D2_C32_Estadao	0,5591133	0,769912
	D3_C32_JB	0,566831683	0,776824
	D4_C32_GPovo	0,561507937	0,795699
	Sumário	0,657894737	0,825581
C33	D2_C33_Estadao	0,47312961	0,708929
	D3_C33_OGlobo	0,566191446	0,793814
	D4_C33_JB	0,579881657	0,78022
	Sumário	0,570945946	0,781609
C34	D1_C34_Folha	0,540342298	0,718487
	D2_C34_Estadao	0,392335766	0,521595
	D3_C34_JB	0,661016949	0,84
	Sumário	0,668639053	0,793814
C35	D1_C35_Folha	0,530612245	0,731801
	D4_C35_JB	0,694736842	0,888889
	D5_C35_GPovo	0,634854772	0,8
	Sumário	0,715277778	0,851852
C36	D1_C36_Folha	0,572815534	0,803371
	D3_C36_OGlobo	0,613793103	0,848485
	D4_C36_JB	0,54460719	0,788636
	Sumário	0,666666667	0,882812
C37	D1_C37_OGlobo	0,624087591	0,835443
	D2_C37_GPovo	0,585492228	0,807692
	Sumário	0,777777778	0,913043
C38	D1_C38_Folha	0,68452381	0,848485
	D2_C38_Estadao	0,589928058	0,732143
	D4_C38_JB	0,666666667	0,862745
	Sumário	0,808219178	0,930233
C39	D2_C39_Estadao	0,57615894	0,780347

	D3_C39_OGlobo	0,577319588	0,747191
	D4_C39_JB	0,550819672	0,780899
	Sumário	0,704081633	0,892857
C40	D2_C40_Estadao	0,602564103	0,82963
	D3_C40_OGlobo	0,501416431	0,695652
	D4_C40_JB	0,673202614	0,870968
	Sumário	0,697916667	0,910714
C41	D2_C41_Estadao	0,476190476	0,606796
	D4_C41_JB	0,619883041	0,826531
	D5_C41_GPovo	0,504716981	0,669291
	Sumário	0,627906977	0,777778
C42	D1_C42_Folha	0,46	0,657371
	D2_C42_Estadao	0,464696223	0,645429
	Sumário	0,621794872	0,795455
C43	D1_C43_Folha	0,471655329	0,706107
	D2_C43_Estadao	0,56640625	0,762821
	D4_C43_JB	0,531418312	0,772036
	Sumário	0,625730994	0,823529
C44	D1_C44_Folha	0,463878327	0,652459
	D2_C44_JB	0,608465608	0,833333
	Sumário	0,651162791	0,821782
C45	D1_C45_Folha	0,539714868	0,754325
	D2_C45_Estadao	0,636871508	0,836538
	D3_C45_OGlobo	0,545454545	0,747748
	Sumário	0,692307692	0,909091
C46	D2_C46_Estadao	0,619565217	0,828025
	D3_C46_OGlobo	0,664974619	0,854545
	D4_C46_JB	0,640625	0,848684
	Sumário	0,758241758	0,94
C47	D3_C47_OGlobo	0,607717042	0,846154
	D4_C47_JB	0,515151515	0,756667
	D5_C47_GPovo	0,515151515	0,756667
	Sumário	0,664556962	0,909091
C48	D1_C48_Folha	0,562363239	0,776557
	D2_C48_Estadao	0,583577713	0,775
	Sumário	0,65248227	0,86747
C49	D1_C49_Folha	0,462441315	0,67382
	D3_C49_OGlobo	0,531468531	0,75625
	D4_C49_JB	0,542553191	0,770701
	Sumário	0,664285714	0,833333
C50	D1_C50_Folha	0,479541735	0,708571
	D3_C50_OGlobo	0,438423645	0,645251
	D4_C50_JB	0,54248366	0,784431
	Sumário	0,641618497	0,858586

Com base na Tabela 2, observa-se que a razão TT obtida por cada um dos *softwares* é diferente. Por exemplo, na coleção 50, o texto D1_C50_Folha possui razão TT de

0,479541735 segundo o SCP e de 0,708571 segundo o CMP. Essa diferença deve-se à utilização de métodos distintos de contagem de *types* e *tokens*. Para o restante do trabalho, considerou-se a razão TT obtida apenas pelo CMP por uma questão de uniformidade, posto que as demais estatísticas também foram obtidas pelo CMP.

Sobre a razão TT, especificamente, apresenta-se, na Tabela 3, a média dessa razão em função de cada fonte jornalística do CSTNews e dos sumários.

Tabela 3: Média da razão TT dos textos-fonte e sumários do CSTNews.

	Jornal	**Razão TT**
Texto-fonte	Jornal do Brasil	0,806593
	Gazeta do Povo	0,795905
	O Globo	0,794260
	O Estado de São Paulo	0,762578
	A Folha de São Paulo	0,746070
Sumário	--	0,864125

Os textos do *Jornal do Brasil* apresentam a razão TT mais alta em 18 das 50 coleções, enquanto os textos de *A Folha de São Paulo* apresentam razão TT mais elevada em apenas 5 coleções (Figura 4).

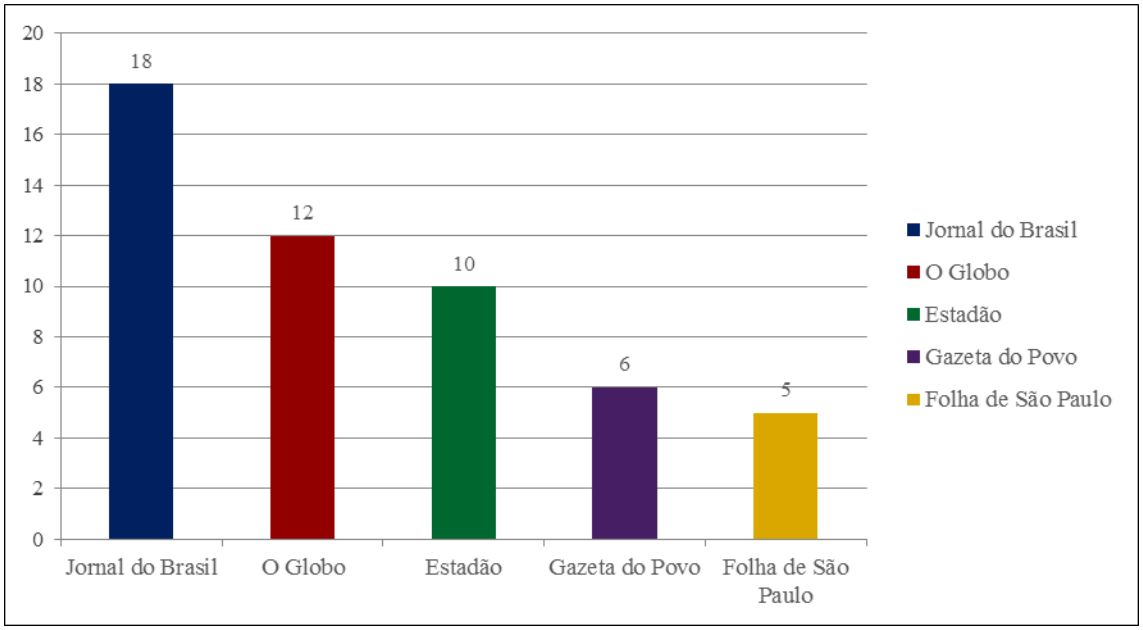

Figura 4: Quantificação dos *clusters* em função dos textos de razão TT mais alta.
Fonte: Elaborada pela autora.

4.2. O tamanho dos textos-fonte e sumários

O cálculo do tamanho dos textos-fonte e dos sumários foi feito por meio do CMP. Especificamente, o CMP calcula o número de palavras a partir do número de *tokens.*

Assim, o tamanho médio dos textos do CSTNews em função das fontes jornalísticas e dos sumários é o apresenta na Tabela 4.

Tabela 4: Tamanho médio dos textos-fonte e sumários do CSTNews.

	Jornal	**Tamanho médio** *(tokens*)
Texto-fonte	A Folha de São Paulo	389,5
	O Estado de São Paulo	346,8
	Gazeta do Povo	340,1
	O Globo	302,2
	Jornal do Brasil	296,8
Sumário	--	137,36

Quanto ao tamanho, os textos do *Jornal do Brasil* possuem em média 296,8 *tokens*, a menor média entre as fontes. Já os textos de *A Folha de São Paulo* possuem tamanho médio de 389,5 *tokens*, ou seja, a maior média entre as fontes. Os sumários em, especial, possuem em média 137,36 *tokens.*

4.3. A incidência de palavras de conteúdo

A métrica "incidência de palavras de conteúdo", que calcula a ocorrência de substantivos, verbos, adjetivos e advérbios, foi obtida por meio do CMP. A média dessa medida em função das fontes está descrita na Tabela 5.

Tabela 5: Incidência média das palavras de conteúdo nos textos-fonte e sumários do CSTNews.

	Jornais	**Incidência das palavras de conteúdo**
Textos-fonte	A Folha de São Paulo	575,7
	Jornal do Brasil	573,3
	Gazeta do Povo	572,7
	O Estado de São Paulo	570,7
	O Globo	567,2
Sumários	--	573,3

Na Tabela 4, vê-se que a maior média foi obtida pelo jornal *A Folha de São Paulo*, ou seja, 575,7070. No entanto, ressalta-se que as médias não são muito discrepantes entre si; talvez a exceção seja a média de *O Globo,* um pouco abaixo das demais (567,2).

4.4. Complexidade/inteligibilidade textual (Índice *Flesch*)

De acordo com as categorias de complexidade textual do "índice *Flesch*" ou Índice de Legibilidade Flesch (ILF), verificou-se que os 140 textos-fonte do *corpus* estão

distribuídos nas categorias “difícil”, “fácil” e “muito difícil” conforme ilustrado na Figura 5.

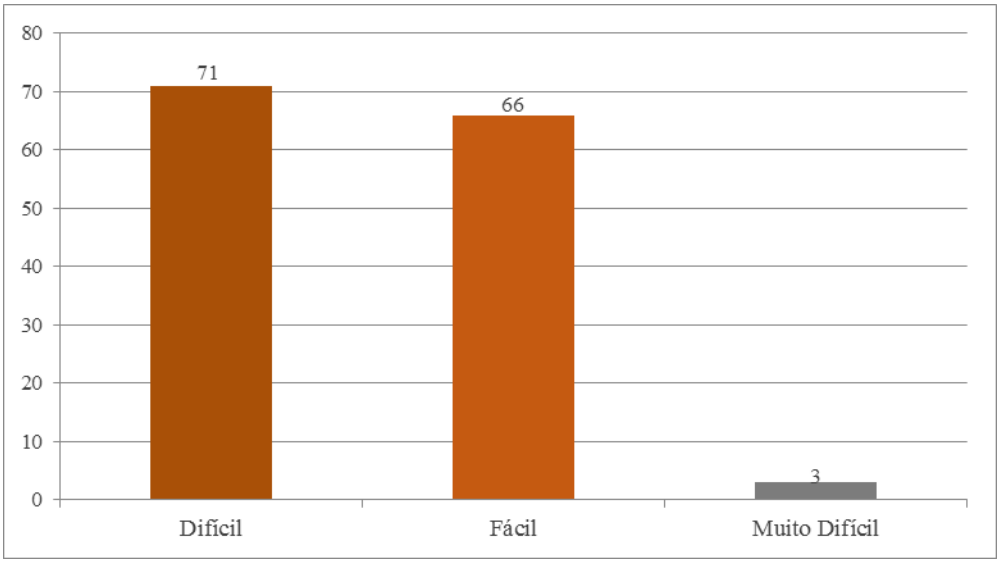

Figura 5: Complexidade textual dos textos-fonte do CSTNews segundo o Índice *Flesch*.
Fonte: Elaborada pela autora.

Tendo em vista as fontes jornalísticas das quais os 140 textos do CSTNews foram compilados, tem-se a distribuição dos textos nas categorias “difícil”, “fácil” e “muito difícil” conforme ilustrado na Tabela 6.

Na Tabela 5, observa-se que: (i) 22 textos do *O Estado de São Paulo* (61%) e 17 do *Jornal do Brasil* (55%) são da categoria difícil, (ii) 17 textos da *Folha de São Paulo* (53%), 13 do *O Globo* (54%) e 10 da *Gazeta do Povo* (59%) se inserem na categoria fácil. A maioria dos textos de o *Jornal do Brasil* e *O Estadão* são da categoria difícil e, por isso, de leitura mais complexa. Já os textos dos jornais *Folha de São Paulo*, *O Globo* e *Gazeta do Povo* são em ampla maioria da categoria fácil, ou seja, de leitura mais acessível.

Tabela 6: Complexidade textual dos textos-fonte do CSTNews em função das fontes jornalísticas.

Fonte	ILF/Categoria	Quantidade de textos	Porcentagem
A Folha de São Paulo	25-50/difícil	15	47%
	50-75/fácil	17	53%
O Estadão de São Paulo	25-50/difícil	22	61%
	50-75/fácil	14	39%
Jornal do Brasil	0-25/muito difícil	2	6%
	25-50/difícil	17	55%
	50-75/fácil	12	39%
O Globo	0-25/muito difícil	1	4%
	25-50/difícil	10	42%
	50-75/fácil	13	54%
Gazeta do Povo	25-50/difícil	7	41%
	50-75/fácil	10	59%

Quanto aos 50 sumários, a maioria deles é da categoria fácil 27 (54%), 22 são da categoria difícil e 1 é da categoria muito difícil, conforme os dados da Tabela 7.

Tabela 7: Complexidade textual dos sumários do CSTNews.

ILF/Categoria	Quantidade de sumários	Porcentagem
0-25/muito difícil	1	2%
25-50/difícil	22	44%
50-75/fácil	27	54%

4.5. Frequência das palavras

Como mencionado, o CMP acessa o Banco de Português, *corpus* de referência do português, e calcula a média de ocorrência de todas as palavras do *texto* no *corpus* de referência. Na Tabela 8, tem-se as médias das frequências em função das fontes e dos sumários. Com na Tabela 8, destaca-se que os textos compilados do jornal *A Folha de São Paulo* possuem, em média, palavras mais frequentes no *corpus* de referência, ou seja, esses textos veiculam palavras que ocorrem bastante no Banco de Português. Ao passo que a Gazeta do Povo, apresenta a menor média.

Tabela 8: Média da "Frequência" em função das fontes do CSTNews.

	Jornais	Média da métrica "Frequência"
Fonte	A Folha de São Paulo	247378,625
	O Estadão de São Paulo	238422,9444
	Jornal do Brasil	229862,3
	O Globo	228311,0417
	Gazeta do Povo	218323,1176
Sumário	--	220191,386

5. Interpretação dos resultados

Diante dos resultados da análise estatística do CMP, tecem-se algumas observações sobre as fontes jornalísticas das quais os textos do CSTNews foram compilados. Para tanto, parte-se da Tabela 9, na qual as médias das diversas medidas estatísticas pesquisadas estão organizadas, sendo que os valores de algumas delas estão destacados em negrito.

Com base na Tabela 9, observa-se, especificamente, que os textos-fonte do *Jornal do Brasil* apresentam a média mais alta da razão TT, isto é, 0,806593. Por esse indicativo estatístico, pode-se dizer esses textos do CSTNews possuem maior riqueza ou diversidade vocabular.

Quanto ao tamanho médio, os textos do *Jornal do Brasil* são em média os menores do *corpus* (296,8 *tokens* em média). A incidência de palavras de conteúdo e a frequência são medidas estatísticas cujas médias não se destacam dos demais.

Sobre o ILF, ressalta-se que uma parcela muito pequena dos textos do *Jornal do Brasil* (2%) é de leitura considerada "muito difícil".

Tabela 9: Médias das diversas estatísticas lexicais referentes aos textos-fonte do CSTNews

	Média das medidas estatísticas					
Fonte	**Razão TT**	**Tamanho**	**Incidência/palavra de conteúdo**	**Frequência**	**ILF**	
A Folha de São Paulo	**0,746**	**389,5**	575,7	247378,6	47%	Difícil
					53%	Fácil
O Estadão de São Paulo	0,762	346,8	570,7	238422,9	61%	Difícil
					39%	Fácil
Jornal do Brasil	**0,806**	**296,8**	573,3	229862,3	2%	**Muito difícil**
					17%	Difícil
					12%	Fácil
O Globo	0,794	302,2	**567,2**	228311,0	1%	**Muito difícil**
					10%	Difícil
					13%	Fácil
Gazeta do Povo	0,795	346,8	572,2	218323,1	7%	Difícil
					10%	Fácil

Ainda quanto a razão TT, vê-se por meio dos resultados que os textos provenientes do jornal *A Folha de São Paulo* apresentam a média mais baixa, isto é, 0,746070. Assim, pode-se dizer que, segundo o parâmetro razão TT, os textos compilados dessa fonte jornalística apresentam menos riqueza ou densidade vocabular. Quanto ao tamanho médio, os textos de *A Folha de São Paulo* são em média os maiores do *corpus* (389,5 *tokens* em média).

Os textos do jornal *O Globo* apresentam a média mais baixa da medida "incidência de palavras de conteúdo" (567,2), sendo que as médias das demais fontes são relativamente similares. Além disso, destaca-se que uma pequena parcela dos textos de *O Globo* é considerada de leitura "muito difícil" (1%).

Na Tabela 10, reuniu-se o conjunto de informações estatísticas sobre as características lexicais dos sumários do CSTNews. Dessa Tabela, destaca-se que uma parcela de apenas 2% dos sumários foi classificada como de complexidade textual "muito difícil". A maior parte dos demais sumários, ou seja, 54%, é considerada de leitura "fácil".

Tabela 10: Médias das diversas estatísticas lexicais referentes aos sumários do CSTNews

	Média das medidas estatísticas					
	Razão TT	**Tamanho**	**Incidência/palavra de conteúdo**	**Frequência**	**ILF**	
Sumário	0,864125	137,36	573,3	220191,386	2%	Muito difícil
					44%	Difícil
					54%	Fácil

6. Considerações Finais

A análise lexical estatística realizada gerou uma série de informações até então pouco exploradas ou mesmo desconhecidas sobre as textos-fonte jornalísticos e os sumários multidocumento do principal *corpus* multidocumento em português, o CSTNews. Consequentemente, conhece-se, hoje, mais características linguísticas lexicais sobre eles do que antes do início do projeto.

Por exemplo, com base na pesquisa realizada, observou-se que os textos compilados de *A Folha de São Paulo* apresentam em média razão TT baixa, ou seja, eles apresentam pouco diversidade vocabular. Isso pode justificar a constatação feita por Carmargo (2013) de que 53% das sentenças dos sumários do CSTNews possuem conteúdo advindo dos textos compilados de *A Folha de São Paulo*. Em outras palavras, levanta-se a hipótese de que a baixa diversidade lexical pode ter influenciado os autores dos sumários multidocumento a ponto de estes se basearem nos textos dessa fonte para a produção dos sumários. No entanto, ressalta-se que a preferência dos autores dos sumários pode ter se baseado em outros critérios.

Além disso, as atividades realizadas durante o projeto permitiram que a aluna conhecesse métricas lexicais e softwares/ambientes de análise estatística com os quais não tinha familiaridade.

Com relação ao cumprimento do cronograma, ressalta-se mais uma vez que a Tarefa 4 (descrição da transposição lexical dos textos-fonte para os sumários) não foi realizada porque as anteriores tomaram mais tempo do que o previsto. Essa demora, aliás, deveu-se principalmente à familiarização da aluna com as métricas e o próprio ambiente do CMP. Pretende-se, no entanto, desenvolver tal tarefa no Trabalho de Conclusão de Curso da aluna, que será desenvolvido no 1º semestre de 2015.

Agradecimentos

Os autores agradecem à FAPESP pelo apoio financeiro.

Referências Bibliográficas

BARBOSA, J.P. **Trabalhando com os gêneros do discurso**: relatar: notícia. São Paulo: FTD, 2001.

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

BICK, E. **The parsing system PALAVRAS**: automatic grammatical analysis of portuguese in a constraint grammar framework. 2000. PhD Thesis. Arhus University, 2000.

CAMARGO, R. T. **Investigação de estratégias de sumarização humana multidocumento**. São Carlos, 2013. 117p. Dissertação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos, São Carlos, 2013.

CARDOSO, P.C.F. et al. A CSTNews - a discourse-annotated corpus for Single and Multi-Document Summarization of news texts in Brazilian Portuguese. In: RST

BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings**… Cuiabá: UFMT, 2011. p.88-105.

DIAS-DA-SILVA, B.C; MONTILHA, G.; RINO, L.H.M.; SPECIA, L.; NUNES, M.G.V.; Oliveira Jr., O.N; MARTINS, R.T; PARDO, T.A.S. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações. **Série de Relatórios do NILC**. NILC-TR-07-10. São Carlos-SP, 119p., 2007.

DOLZ, J.; SCHNEUWLY, B. **Gêneros orais e escritos na escola**. Campinas, SP: Mercado de Letras, 2004. 278 p. (Tradução e organização: Roxane Rojo; Glaís Sales Cordeiro).

LAGE, N. **Estrutura da Notícia**. 5ª ed. São Paulo: Ática, 2002.

LI, P.; WANG, Y.; GAO, W.; JIANG, J. Generating Aspect-oriented Multi-Document Summarization with event-aspect model, ACL WORKSHOP ON AUTOMATIC SUMMARIZATION FOR DIFFERENT GENRES, MEDIA, AND LANGUAGES, 2011, Portland, USA. **Proceedings**… Portland, 2011, p. 1137–1146.

MANI, I. **Automatic Summarization**. Amsterdam: John Benjamins Publishing Co., 2001.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical Natural Language Processing**. Cambridge, MA: MIT Press, 1999.

MAZIERO, E.G. **Identificação automática de relações multidocumento**. São Carlos, 2012, 117p. Dissertação (Mestrado em Ciência da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos, 2012.

RATNAPARKHI, A. A maximum entropy part-of-speech tagger. In: EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING CONFERENCE, 1, 1996. **Proceedings**… Philadelphia, 1996. p. 133-142.

NENKOVA, A. **Understanding the process of multi-document summarization**: content selection, rewrite and evaluation. PhD Thesis, Columbia University, January 2006.

OWCZARZAK, K.; DANG, H.T. Who wrote what where: analyzing the content of human and automatic summaries. ACL WORKSHOP ON AUTOMATIC SUMMARIZATION FOR DIFFERENT GENRES, MEDIA, AND LANGUAGES, 2011, Portland, USA. **Proceedings**… Portland, 2011, p. 25-32.

RADEV, D.R. A common theory of information fusion from multiple text sources, step one: cross-document structure". In: ACL Signal Workshop on Discourse and Dialogue, 1, 2000, Hong Kong, **Proceedings**… Hong Kong, 2000, p. 74-83.

RASSI, A, P.; ZACARIAS, A.C.I.; MAZIERO, E.G.; SOUZA, J.W.C.; DIAS, M.S.; CASTRO JORGE, M.L.R.; CARDOSO, P.C.F.; BALAGE FILHO, P.P.; CAMARGO, R.T.; AGOSTINI, V.; DI-FELIPPO, A.; SENO, E.R.M.; RINO, L.H.M.; PARDO, T.A.S. Anotação de aspectos textuais em sumários do corpus CSTNews. **Série de Relatório Técnico do NILC**, NILC-TR-13-01. São Carlos-SP, Junho, 2013, 55p.

SCARTON, C.E.; ALUÍSIO, S. M. Análise da inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-

Metrix para o Português. **LinguaMÁTICA**, v. 2, n. 1, pág. 45-62, Abr. 2010. ISSN 1647-0818

SCOTT, M. **WordSmith Tools Help**. Liverpool: Lexical Analysis Software. 2014. Disponível em: http://www.lexically.net/downloads/version6/HTML/proc_tag_handling.htm. Acesso em 10 set. 2013.

SPARCK-JONES, K., WILLET, P. **Readings in information retrieval**. São Francisco: Morgan Kaufmann, 1997.

SPARCK-JONES, K. **Discourse modeling for Automatic Summarization**. Tech. Report No. 290. University of Cambridge. UK, February 1993.

SWALES, J. M. (1990). **Genre Analysis**: English in Academic and Research Settings. Cambridge, UK: Cambridge University Press.

VOUTILAINEN, A. Part-of-speech tagging. In: MITKOV, R. (Ed.). **The Oxford handbook of Computational Linguistics**. Oxford, New York: Oxford University Express, 2004, cap. 11, p. 219-232.