Lexicologia e informação: um ensaio de quantificação

Lexicology and information: an essay of quantification

César Nardelli Cambraia*

RESUMO: Este trabalho tem como objetivo avaliar a técnica de medir informação de um conjunto de textos a partir de seu vocabulário. Do ponto de vista teórico, adotam-se conceitos da lexicologia estrutural (BIDERMAN, 2001) e do funcionalismo (NEVES, 1997). Fez-se uma coleta de todos os textos publicados sobre um determinado tema na versão digital de três periódicos de Minas Gerais por nove meses (179 textos) e em seguida fez-se uma análise 22quantitativa (baseada em ocorrências de lexias, lexias diferentes e lexemas diferentes) e uma análise qualitativa (baseada na relevância de certos eventos). Como resultado, verificou-se que: (a) os três periódicos se comportaram diferentemente no tratamento do tema, (b) o nível dos lexemas é o mais relevante para quantificar diferença no volume de informação de um texto, (c) a quantidade de informação nova em cada texto analisado é relativamente baixa (13-15%), (d) os periódicos continuam compondo textos digitais segundo a lógica da mídia impressa, e (e) a quantificação de informação por lexema não se mostrou como critério suficiente para identificação de informação relevante.

PALAVRAS-CHAVE: Lexicologia. Imprensa. Linguística de corpus.

ABSTRACT: This work aims to evaluate the technique of measuring information of a set of texts through its vocabulary. From a theoretical point of view, we adopted concepts of structural lexicology (BIDERMAN, 2001) and functionalism (NEVES, 1997). We collected all texts published about a certain topic on digital version from three journals of Minas Gerais throughout nine months (179 texts) and then we did a quantitative analysis (based on tokens, types and lemmas) and a qualitative analysis (based on the relevance of certain events). As results, we found that: (a) the three journals behaved differently in the treatment of the topic, (b) the level of lemmas is most relevant to quantify differences in the amount of information of a text, (c) the amount of new information in each text analyzed is relatively low (13-15%), (d) periodicals continue to compose digital texts following the logic of printed media, and (e) the quantification of information through lemmas is not a sufficient criterion for identifying relevant information.

*Doutor em Filologia e Língua Portuguesa pela Universidade de São Paulo. Atualmente é professor associado de Filologia Românica na Universidade Federal de Minas Gerais e tem experiência na área de Linguística Românica e Crítica Textual, atuando principalmente nos seguintes temas: estudo histórico e comparado de morfossintaxe de línguas românicas em uma perspectiva tipológico-funcional, lexicologia sócio-histórica e edição de textos românicos antigos. Bolsista de Produtividade em Pesquisa do CNPq - Nível 2. E-mail: nardelli@ufmg.br.

KEYWORDS: Lexicology. Press. Corpus linguistics.

Introdução

O desenvolvimento da mídia digital tem permitido, em função do baixo custo, a produção e a divulgação de um volume sem precedentes de textos. Essa profusão de textos pode apresentar consequências negativas: como identificar presença de informação nova em um mar de textos *on-line*? Quanto maior o volume de textos disponíveis, mais tempo o leitor leva para ter acesso às informações novas, diminuindo assim a eficiência do acesso ilimitado à informação.

Em função dessa situação, desponta para a linguística atual um desafio de especial interesse: considerando que os textos são constituídos no mais das vezes por palavras (embora possam aparecer acompanhados também de imagens), um tratamento automatizado desse substrato poderia oferecer instrumentos que muniriam de maneira adequada o leitor para sua "batalha" na busca de informação nova. Tem-se, portanto, uma circunstância especialmente interessante para aliar os recursos da informática com um instrumental teórico da linguística, em particular, da lexicologia.

No presente trabalho, apresenta-se um estudo-piloto que teve como objetivo avaliar uma técnica para medir informação de um conjunto de textos a partir de seu vocabulário.

Lexicologia e Informação

A lexicologia é o ramo da linguística que se ocupa do léxico de uma língua. O léxico, por sua vez, consiste no conjunto de palavras de uma dada língua. Considerando que o termo *palavra* é utilizado com várias acepções, convém ser mais preciso adotando-se os termos *lexema* e *lexia*. Biderman (2001, p. 169) define *lexema* como a unidade léxica abstrata da língua e *lexia*

como a forma que aparece no discurso. Tradicionalmente, representa-se metalinguisticamente o primeiro com maiúsculas (ou versalete) e o segundo com itálico. Um lexema é um paradigma que abarca todas as formas flexionadas: por exemplo, o lexema verbal VER abarca formas concretas como *ver, vendo, visto, vejo, vias, viu, veremos, vísseis, viram*, etc. Entretanto, quando a oposição entre as formas se dá em termos derivacionais, tem-se diferentes lexemas, como em análise, an22alisar, analisador, analisável, etc.

Costuma-se dividir os lexemas em duas grandes categorias: *de conteúdo* e *instrumentais* (BIDERMAN, 2001, p. 333). Os primeiros, também chamados de *nocionais*, se caracterizam por ter significação externa, ou seja, por representarem referentes do mundo extralinguístico; já os segundos, também conhecidos por *gramaticais*, desempenham sobretudo a função de estabelecer relações entre os de conteúdo. Na teoria, a divisão parece sustentável, mas, na prática, o problema é mais complexo: o exemplo mais evidente está na classe das preposições, em que certos itens poderiam desempenhar tanto função referencial (como DE em frases como *Venho de Belo Horizonte*, em que o valor nocional de ponto de origem está necessariamente vinculado ao DE) quanto função relacional (como em frases como *Gosto de você*, em que a preposição DE tem como única função marcar sintaticamente o complemento do verbo GOSTAR)[1].

A identificação de unidades lexicais na cadeia da fala segundo critérios propriamente linguísticos está longe de ser pacífica: Biderman (2001, p. 137-155), por exemplo, assinala a necessidade de se combinarem critérios de natureza fonológica, morfossintática e semântica. Entretanto, essa mesma autora reconhece a imposição operacional de se considerar o critério gráfico para análises de *corpora* extensos: assim o fez em sua pesquisa para um dicionário de frequência do português brasileiro contemporâneo (BIDERMAN, 1998).

[1] Para uma discussão sobre essa problemática especificamente no caso das preposições, cf. Berg (2005).

De todos os níveis de organização da estrutura linguística, o léxico é certamente o de maior interesse para quantificação de informação, dado principalmente o valor referencial dos lexemas nocionais[2]. Isso não significa, no entanto, que outros níveis de organização da linguagem não veiculem também informação: basta lembrar aqui aspectos como o deslocamento de constituintes sintáticos com o objetivo de marcar funções como tópico, foco, etc. Como estratégia de quantificação de informação de forma automatizada por meio de recursos da informática, seguramente o léxico ocupa um lugar privilegiado, dada a facilidade de se coletarem automaticamente suas unidades.

Modelos teóricos da linguística de orientação funcionalista têm insistido na iconicidade da linguagem humana, ou seja, na existência de "uma relação não-arbitrária entre forma e função, ou entre código e mensagem" (NEVES, 1997, p. 103).

Um dos princípios que regeriam a organização da linguagem humana seria o de *isomorfia*, ou seja, "uma forma para um significado e um significado para uma forma" (BOLINGER, 1977 apud NEVES, 1997, p. 105). Assim, por exemplo, o lexema COMPLACÊNCIA e sua proposição definitória, como a de que seja uma "disposição habitual ou tendência de corresponder aos desejos, gostos, idiossincrasias de outrem com a intenção de ser-lhe agradável" (HOUAISS et al., 2001), não apresentariam identidade semântica: seriam apenas aproximações. A discussão sobre a representação do significado é bastante complexa e não será abordada aqui, mas convém ressaltar que reflexões atuais tem salientado a necessidade de se repensar a noção de significado, superando-se uma visão referencialista (significado como representação do mundo), em prol de uma visão sociocognitivista (significado como ação discursiva sobre o mundo), como já assinalou Marcuschi (2004).

Um segundo princípio defendido pelos funcionalistas é o *princípio da quantidade*: "um texto maior deve conter mais informação do que um texto

[2] Dado o vínculo entre mudanças no léxico e transformações sociais, esse nível linguístico se destaca também como objeto de interesse para abordagens sócio-históricas da língua (CAMBRAIA, 2013; MATORÉ, 1973).

menor, já que, admitindo-se a relação icônica entre forma e organização do conteúdo, maior quantidade de matéria fônica deve corresponder a maior quantidade de informação" (NEVES, 1997, p. 107). Considerando que um lexema pode ter vários significados, poder-se-ia imaginar que um único lexema conteria mais informação do que, por exemplo, o conjunto de suas proposições definitórias (HOUAISS et al., 2001), p. ex., apresentam sete para o lexema COMPLACÊNCIA). Essa aparente contradição, no entanto, é superada ao se considerar que cada item presente na própria proposição definitória apresenta igualmente uma multiplicidade de significados, fazendo com que de fato, as proposições definitórias sempre reúnam em si uma gama maior de conteúdo do que o veiculado apenas pelo lexema que definem.

A discussão sobre os princípios da iconicidade, da isomorfia e da quantidade foi realizada aqui para se sustentar a ideia de que a extensão de um texto é sim indício do volume de informação que contém. Embora não seja a única estratégia possível de se medir volume de informação, não se pode negar que seja de fato uma estratégia relevante.

Aceitando-se o postulado de que a extensão de um texto é indício do volume de informação que contém, coloca-se de imediato a questão sobre qual é o critério que se deve empregar para essa medida. Poder-se-ia pensar em número de parágrafos ou de frases, mas se trata de critérios por demais limitados, já que a extensão dos parágrafos e das frases pode variar sensivelmente. Passando do nível de parágrafos e de frases para um nível de elementos com extensão menos variável, tem-se como elemento relevante a palavra. Mas é preciso avaliar com que nível de organização das palavras dever-se-ia trabalhar: a medida da extensão pode ser feita, por exemplo, levando-se em conta o número de ocorrências de lexias de cada texto, o número de lexias diferentes ou o número de lexemas diferentes. A ausência de estudo que tenha demonstrado qual desses níveis de análise seria o mais apropriado para quantificação do volume de informação sugere que se deva, neste estudo, avaliar os três.

A identificação de estratégias interessantes para quantificar informação, a partir de análises lexicais, permitirá, por exemplo, a construção de algoritmos que realizem, de forma automática, a análise de uma grande quantidade de textos sobre um mesmo tema para se identificar quais deles possuem um maior volume de informação, fornecendo assim instrumentos para os leitores navegarem de forma mais produtiva no incomensurável mar de textos do mundo digital. Não se trata obviamente de uma questão totalmente nova, uma vez que buscadores (como o *Google*) já utilizam algoritmos para gerar seus resultados: a novidade está em avaliar diferentes formas de se analisarem textos com base no seu vocabulário.

Um Estudo de Caso

Partindo do postulado de que a extensão de um texto é indício do volume de informação que contém**,** utiliza-se aqui este critério para analisar um conjunto de textos sobre um mesmo tema produzidos por três diferentes periódicos diários publicados em Belo Horizonte, tomado como referência seu vocabulário. O objetivo deste estudo é avaliar essa técnica para medir informação de um conjunto de textos a partir de seu vocabulário.

Metodologia

Como tema para o presente estudo, elegeu-se a morte dos dois jornalistas mineiros em 2013 (fatos relacionados entre si) que atuavam no Vale do Aço, na região leste de Minas Gerais. A escolha desse tema justifica-se por duas razões: (a) temas de caráter policial costumam receber especial atenção na mídia, dando origem a um volume maior de reportagens do que temas com menor apelo popular; e (b) dada a complexidade dos eventos, houve nesse caso específico, vários pontos de virada no curso da investigação, gerando assim momentos de pico de informação nova.

Escolheram-se como periódicos para coleta de dados os três jornais com maior difusão em Minas Gerais: *O Tempo* (OT), *Estado de Minas* (EM) e *Hoje em Dia* (HD). O interesse em trabalhar com mais de um periódico deriva do desejo de avaliar também se os veículos da imprensa se comportam de forma diferente em relação ao tratamento de um mesmo tema.

Primeiramente, buscaram-se na base de dados dos três periódicos notícias com as expressões "R. N." (nome do primeiro jornalista morto) e "W." (nome do segundo jornalista),[3] processo que poderia de fato ser feito de forma puramente automática; mas, através da leitura de cada notícia retornada na busca, identificaram-se também as notícias correlatas, mesmo que não estivessem no resultado da busca feita inicialmente, procedimento que uma busca automática teria dificuldade para realizar.

Especificamente no caso de OT, as buscas retornaram notícias tanto da publicação principal quando da de caráter mais popular (jornal *Super*): nesse caso, foram consideradas apenas as notícias da publicação principal, porque em muitos casos o texto de ambas era praticamente o mesmo, o que poderia viciar a análise.

Todas as notícias foram coletadas e salvas em arquivos *txt* (formato necessário para o processamento pelo software de análise que será adiante informado). De cada página *html* do portal de cada periódico, extraiu-se apenas o texto da notícia, incluído o tema (geralmente antes do título), título, subtítulo, dados de publicação (data e hora), autor (quando havia) e texto propriamente dito. As chamadas com *links* para outras notícias (geralmente com seu título) não foram coletadas. Nos casos em que houve imagem, sua legenda e seu texto (quando houvesse) foram incluídos como parte do artigo.

A análise de cada texto foi feita basicamente em três níveis: (a) número de ocorrências de lexias (ing. *tokens*); (b) número de lexias diferentes (ing. *types*); e (c) número de lexemas diferentes (ing. *lemmas*). Para a quantificação

[3] A busca foi feita com base no nome por extenso dos jornalistas, mas aqui, por razões éticas, são apresentadas apenas as iniciais. A forma por extenso dos nomes foi específica o suficiente para evitar resultados com dados não relacionados ao tema. Em todos os textos considerados na análise há a presença do nome dos dois jornalistas ou de pelo menos um deles.

desses valores, utilizou-se o software *Wordsmith* (versão 6.0.0.166, de 07/12/2013). Especificamente para a quantificação de lexemas diferentes, foi necessário construir uma lista de lexemas, associando as diferentes lexias a seu respectivo lexema. A construção dessa lista foi feita de forma semiautomática: a partir da lista de lexias gerada pelo *Wordsmith* com base em todos os textos do *corpus*, fez-se primeiramente uma conversão automática pela substituição de morfemas flexionais nominais e verbais (p. ex., *-ndo* > -R, como em *faze**ndo*** > FAZE**R**); posteriormente, realizou-se uma revisão manual corrigindo-se as distorções (como em *Ferna**ndo*** > FERNAR em vez de FERNANDO). O método de conversão automática usado apresentou uma taxa de 70% de acerto: os 30% de erros na lista de lexemas foram corrigidos manualmente.

Resultados

Pela coleta de dados segundo o método descrito na seção anterior, localizaram-se no total 179 textos publicados nos últimos nove meses (de 8 de março a 7 de dezembro de 2013), assim distribuídos por periódico:

Tabela 1: Textos de notícia por mês[4]

	OT	**EM**	**HD**
08/03-07/04	14	8	7
08/04-07/05	36	19	17
08/05-07/06	20	5	7
08/06-07/07	4	4	4
08/07-07/08	5	6	2
08/08-07/09	3	4	2
08/09-07/10	1	1	0
08/10-07/11	3	2	1
08/11-07/12	1	2	1
Total	**87**	**51**	**41**

Fonte: O autor.

[4] Foram coletados apenas os textos de notícias disponíveis para o público não-assinante. Em apenas um caso (EM, 13/11/2013), excluiu-se parte do texto da página digital, pois havia duas notícias independentes na mesma página, tendo-se ficado apenas com a relativa ao tema em estudo.

Pelos dados da Tabela 1, percebe-se claramente que a atenção dada ao tema não é a mesma entre os periódicos. No que se refere a número de textos, OT (87) apresenta no conjunto quase o dobro de textos do que EM (51) e HD (41) apresentam.

A análise lexical foi realizada levando-se em conta o número de ocorrências de lexias (L1), o de lexias diferentes (L2) e o de lexemas diferentes (L3)[5]. Convém esclarecer que o total na base das colunas L2 e L3 não é a soma simples dos totais por mês, mas sim o total no conjunto dos textos de um mesmo periódico (formas que apareceram em mais de um mês foram computadas apena uma vez para o total final).

Tabela 2: Número total de ocorrências de lexias, de lexias diferentes e de lexemas diferentes por mês[6]

	OT			**EM**			**HD**		
	L1	**L2**	**L3**	**L1**	**L2**	**L3**	**L1**	**L2**	**L3**
08/03-07/04	3.516	1.977	585	2.976	1.600	539	2.830	1.512	569
08/04-07/05	10.232	5.598	1.444	13.214	5.913	1349	6.319	3.311	846
08/05-07/06	6.172	3.344	820	2.591	1.227	511	3.226	1.681	646
08/06-07/07	1.073	582	254	1.688	845	301	1.482	801	366
08/07-07/08	1.572	842	433	5.201	2.269	824	1.222	529	361
08/08-07/09	5.207	1.665	1.046	2.055	967	514	804	422	314
08/09-07/10	339	172	154	421	212	189	0	0	0
08/10-07/11	1.128	574	356	1.243	586	445	607	300	262
08/11-07/12	292	155	141	522	267	163	414	220	196
Total	**29.531**	**3.343**	**2.237**	**29.911**	**3.147**	**2.040**	**16.904**	**2.189**	**1.504**

Fonte: O autor.

[5] Não se aplicou nenhuma lista de exclusão no processamento: os dados incluem, portanto, lexemas nocionais (substantivos, adjetivos, verbos e advérbios em –mente) e gramaticais (demais advérbios, artigos, preposições, conjunções, interjeições, numerais e pronomes).

[6] No recurso *Wordlist* do programa *Wordsmith*, a categoria L1 da Tabela 2 corresponde à coluna *tokens (running words) in text*; L2, à *types (distinct words)*; e L3, ao número de palavras depois de aplicado o recurso de *zapping* (supressão das lexias subordinadas a lexemas segundo a lista de lexemas inserida no processo).

Primeiramente é interessante notar que o número de ocorrências de lexias (L1) não repete a hierarquia de número de textos vista na Tabela 1 (OT > EM > HD), pois agora é EM > OT > HD. É curioso perceber que os 87 textos de OT contêm um número de ocorrências de lexias menor que os 51 de EM, estando aquele 1% atrás deste nesse aspecto. No que se refere ao número de lexias diferentes (L2), a hierarquia da Tabela 1 (OT > EM > HD) volta a proceder, estando OT 6% à frente de EM. Por fim, quanto ao número de lexemas diferentes (L3), percebe-se que a referida hierarquia (OT > EM > HD) também se mantém, estando agora OT 10% à frente de EM nessa questão.

A diferença entre o 1% no número de ocorrências de lexias e os 10% no de lexemas diferentes entre OT e EM evidencia que a análise do número de lexemas diferentes (L3) parece ser mais proveitosa para identificar diferenças do que a de níveis como número de ocorrências de lexias (L1) ou número de lexias diferentes (L2).

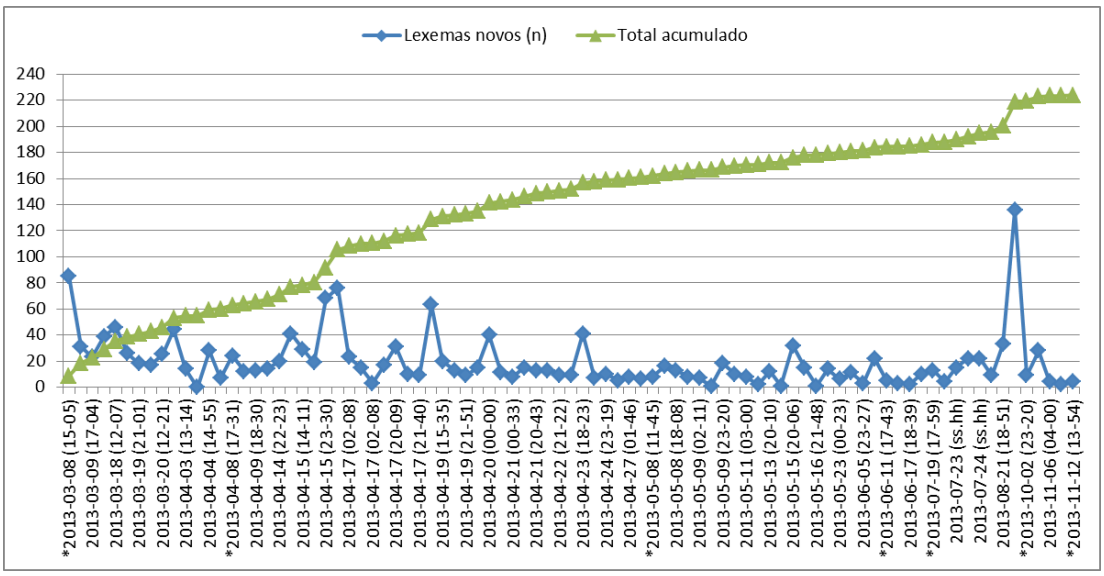
Os dados das Tabelas 1 e 2 confirmam, portanto, a suposição de que as mídias se comportam diferentemente no tratamento de uma mesma notícia: há diferença tanto no número de textos (sobretudo de OT frente a EM e HD) quanto no de lexemas diferentes (sobretudo de OT e EM frente a HD).

Dado o fato de que os textos coletados foram produzidos ao longo de nove meses, parece interessante avaliar como foi a progressão de informação nova. Esse aspecto pode ser avaliado calculando-se o número de lexemas novos a cada novo texto publicado pelos veículos da imprensa, considerando-se os lexemas do conjunto de textos já publicados até então. Deve-se, porém, assinalar que os lexemas empregados, até então normalmente, não estão disponíveis ao leitor em cada novo texto, porque a cada novo texto há a introdução de lexemas novos (equivalendo à informação nova) mas há também a supressão de lexemas já empregados (equivalendo à informação velha).

Nos gráficos abaixo apresentam-se o número de lexemas novos (n) e o número total acumulado de lexemas (para este último, utilizou-se o logaritmo na base de 10 para ser melhor visualizado em um mesmo gráfico com o

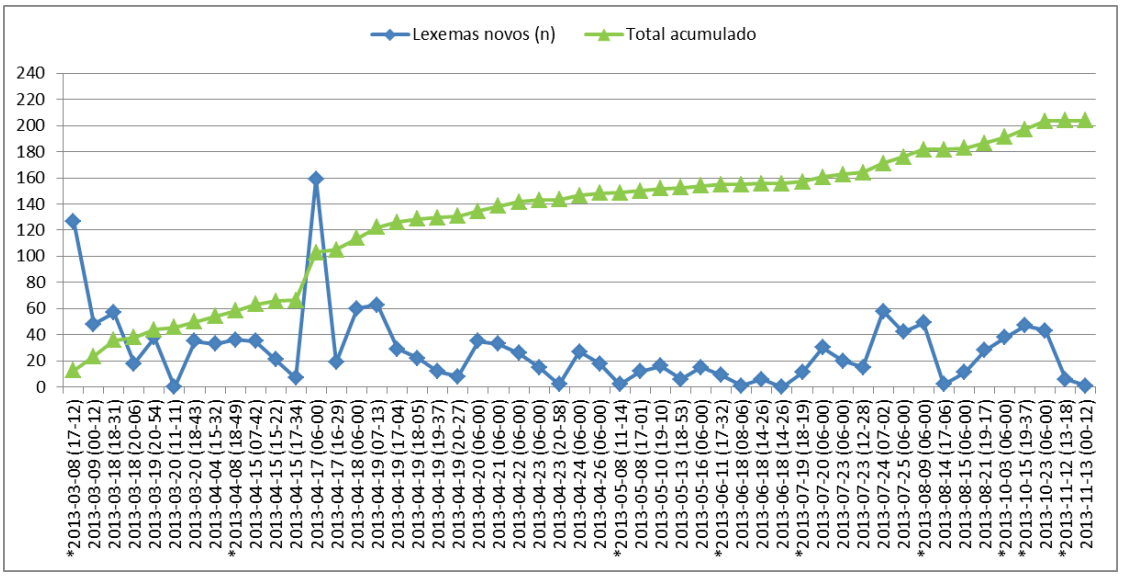
número de lexemas novos). Informa-se no eixo horizontal a data de publicação do texto com a indicação da hora entre parênteses e no eixo vertical o número de lexemas novos.

Gráfico 1 - Lexemas novos e total acumulado (OT)[7]



Fonte: O autor.

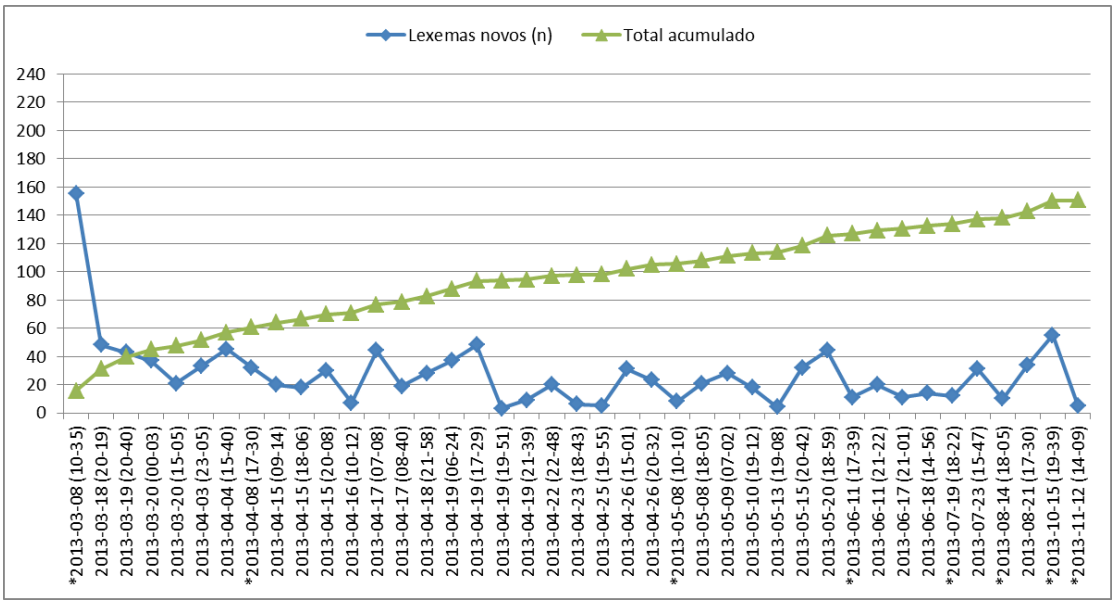
Gráfico 2 - Lexemas novos e total acumulado (EM)



Fonte: O autor.

[7] O asterisco antes da data no gráfico indica início de nova faixa mensal.

Gráfico 3 - Lexemas novos e total acumulado (HD)



Fonte: O autor.

Uma análise comparativa entre os Gráficos 1, 2 e 3 permite verificar, primeiramente, um padrão ondular no número de lexemas novos, mais nítido em EM e HD: aceitando-se o número de lexemas novos como indício de informação nova, tem-se que o fluxo de informação nova se dá de forma variável, alternado momentos de aumento com de diminuição.

Esse padrão ondular sugere que, mesmo não havendo muita informação nova, os periódicos consideram relevante publicar um texto sobre o tema. Esse aspecto é especialmente nítido no caso de OT, com seus 87 textos: justamente porque são tantos, o volume de informação nova (expressa por lexemas novos) a cada texto é bastante baixo, como se constata pela sua média (19 lexemas novos por texto), frente à de EM (28) e de HD (27). Pode-se imaginar que essa prática de publicação de novos textos sobre um tema, quase que diariamente, mesmo não tendo tanta informação nova seja, na verdade, uma afirmação política frente ao tema: torná-lo visível todo dia seria uma estratégia de mostrar a importância que se dá ao tema (no caso, a apuração da morte de profissionais da imprensa). A natureza política dessa prática de reiteração de informação velha é sugerida sobretudo pela existência de textos em que o número de lexemas novos é próximo ou igual a zero, ou seja, não haveria

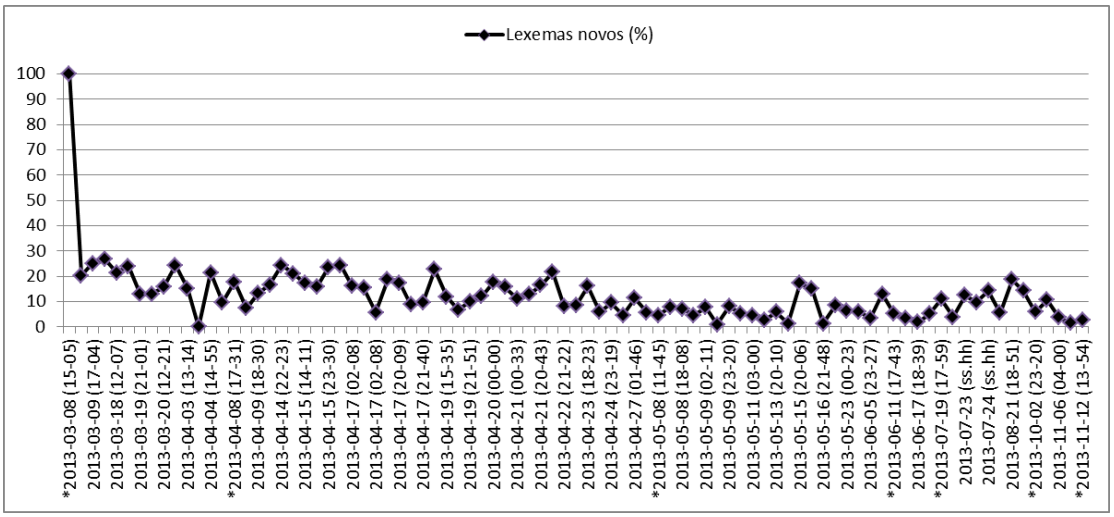
nenhuma informação nova nesses novos textos, como no caso de OT (03/04/2013, 23:27) e EM (20/03/2013, 11:11) com nenhum lexema novo, e de HD (19/04/13, 19:51), com apenas três.

Poder-se-ia argumentar, porém, que a reiteração de informação velha seria uma decorrência na natureza do texto: como os artigos de jornais não são lidos necessariamente todos os dias pelas mesmas pessoas, então seria sempre preciso incluir informação velha (pelo menos uma breve síntese) ao incluir cada informação nova, para que o leitor que entrasse em contato pela primeira vez com o tema tivesse um mínimo de contextualização. Sob essa perspectiva, compreende-se a repetição de informação velha a cada novo texto, mas certamente gera-se uma sobrecarga sobre o leitor que acompanhe diariamente o tema: para este, haveria sempre o fardo de passar por muita informação velha para se ter acesso à informação nova.

Quiçá a mídia atual, agora em formato digital, ainda não tenha conseguido se desligar de uma concepção segundo a qual um leitor não tem acesso às edições anteriores, o que certamente seria verdade, de forma geral, na época em que os periódicos circulavam apenas em versão impressa, pois ninguém tinha o hábito de guardar todas as edições para sempre. Tem-se hoje uma situação mista, na qual os textos de notícia ainda apresentam na sua configuração padrões próprios de uma situação em que o leitor não teria mais acesso a textos anteriores, mesmo essa situação já tendo ficado para trás, uma vez que, como já mencionado, hoje em cada página *html* com o texto de notícia há vários *links* chamando o leitor para textos anteriores já publicados sobre o tema (saliente-se, porém, que não parece haver clareza em relação ao critério usado por cada periódico para selecionar esses *links* tanto qualitativamente quanto quantitativamente). Essa situação mista parece decorrer do fato de que as edições impressas e digitais partilham muitos textos: um redator escreve, normalmente, uma só versão que possa circular tanto na edição impressa quanto na digital, mesmo tendo cada uma dessas mídias uma lógica de organização intrinsecamente distinta.

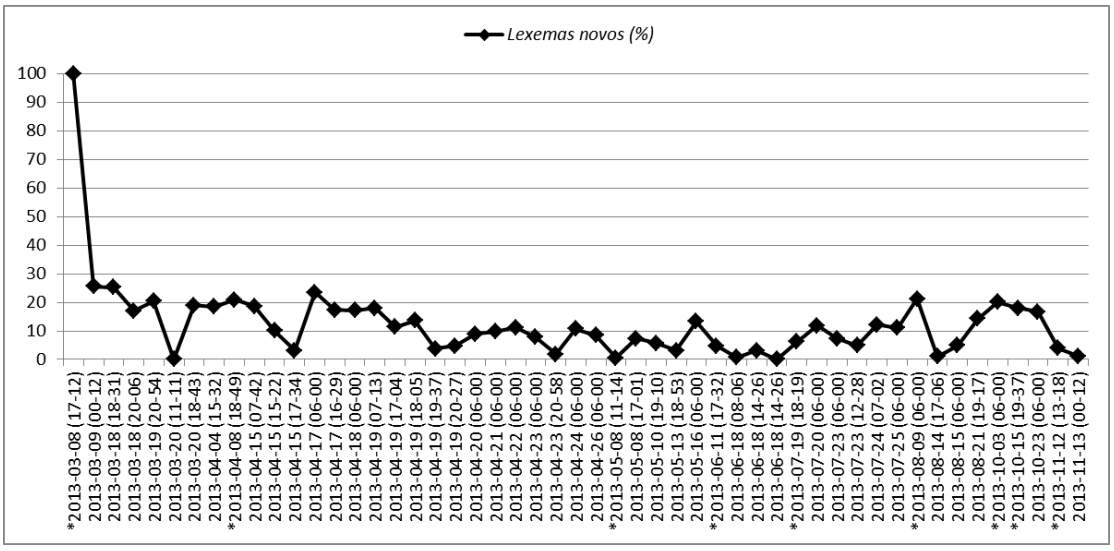
Outra medida para avaliar o volume de informação nova através da contagem de lexemas pode ser obtida por meio da porcentagem de lexemas novos em relação ao lexemas totais em cada texto. Essa medida pode oferecer mais dados, pois se estará avaliando não apenas se há informação nova, mas também qual é sua porcentagem em relação ao texto como um todo.

Gráfico 4 - Lexemas novos (OT)



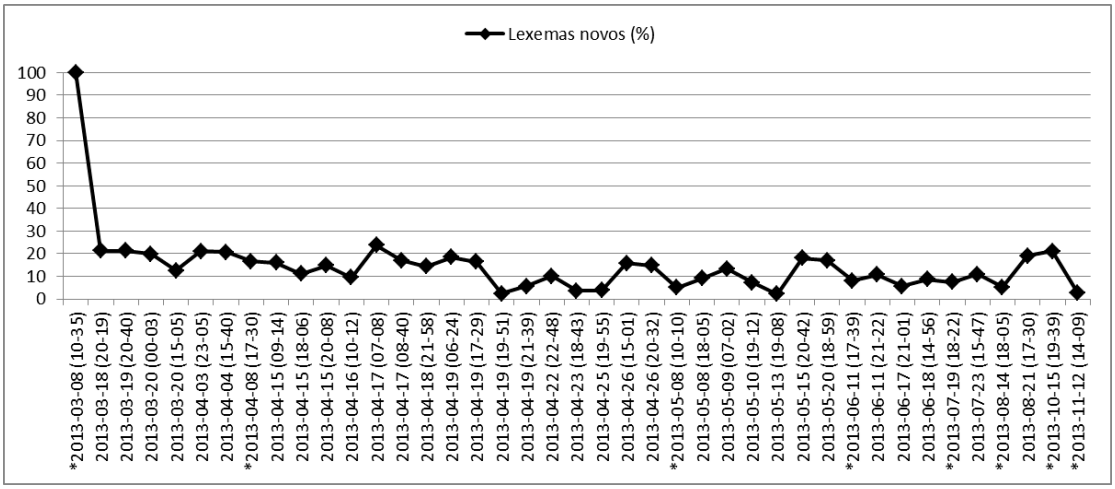
Fonte: O autor.

Gráfico 5 - Lexemas novos (EM)



Fonte: O autor.

Gráfico 6 - Lexemas novos (HD)



Fonte: O autor.

É interessante notar que a porcentagem de lexema novo (e, portanto, de informação nova) é bastante baixa a cada texto: as médias são OT (13%), EM (13%) e HD (15%) – há, aliás, praticamente empate técnico. Comparando os Gráficos 1, 2 e 3 respectivamente com os Gráficos 4, 5 e 6, percebe-se que o padrão ondular de informação nova mantém-se. Essa correlação deve derivar de uma imposição de ordem prática à estruturação de uma notícia: ela precisa ter uma dimensão mais ou menos fixa, ou seja, cada texto jornalístico teria um valor mais ou menos fixo de lexemas. As médias de lexemas totais por textos são OT (148), EM (229) e HD (185) – vê-se que, mesmo havendo um limite básico para o tamanho da notícia, ele não é o mesmo para cada jornal. Por um lado, percebe-se em OT a relação entre mais textos (87) e menos lexemas (148) em relação aos demais; por outro lado, a relação inversa não é nítida nos dois outros, pois o periódico com menos textos (HD, 41) não é o com mais lexemas: o valor mais alto está com EM (229).

Em síntese, o baixo número de lexemas novos a cada novo texto dos periódicos decorreria da necessidade de sempre se contextualizar a informação nova com informação velha; já a baixa porcentagem de lexemas novos a cada novo texto derivaria da obrigação de se respeitar um limite-padrão para a dimensão dos textos. Saliente-se que o baixo número de lexemas novos não necessariamente deveria acarretar uma baixa porcentagem de lexemas novos,

pois uma diminuição progressiva de informação velha a cada novo texto poderia levar a um aumento significativo da porcentagem de informação nova, mas isso não acontece porque os periódicos mantêm uma prática de produzir textos com uma dimensão mais ou menos padronizada, independentemente do volume de informação nova.

Por fim, convém ensaiar um contraste entre as análises quantitativas realizadas até aqui e uma análise propriamente qualitativa. Uma estratégia para isso seria a identificação de momentos críticos dos eventos (baseados na divulgação de fatos essenciais para a elucidação da questão) e sua comparação com o fluxo de lexemas novos quantificado acima. Essa estratégia tem como limitação a dificuldade de se estabelecer de forma totalmente objetiva quais fatos são os mais importantes em relação ao tema. Para minorar esse problema, será feita a escolha de apenas 10 fatos importantes. Como apoio para a identificação desses fatos, usaram-se especialmente os textos de notícias com retrospectivas sintéticas dos fatos, como, p. ex., OT (15/04/13, 08:14). No curso dos eventos de 08 de março de 2013 (data do assassinato de R.N.) a 07 de dezembro desse mesmo ano (data-limite final da coleta de dados), teriam sido especialmente importantes os seguintes fatos:

Quadro 1 - Seleção de 10 momentos críticos dos eventos

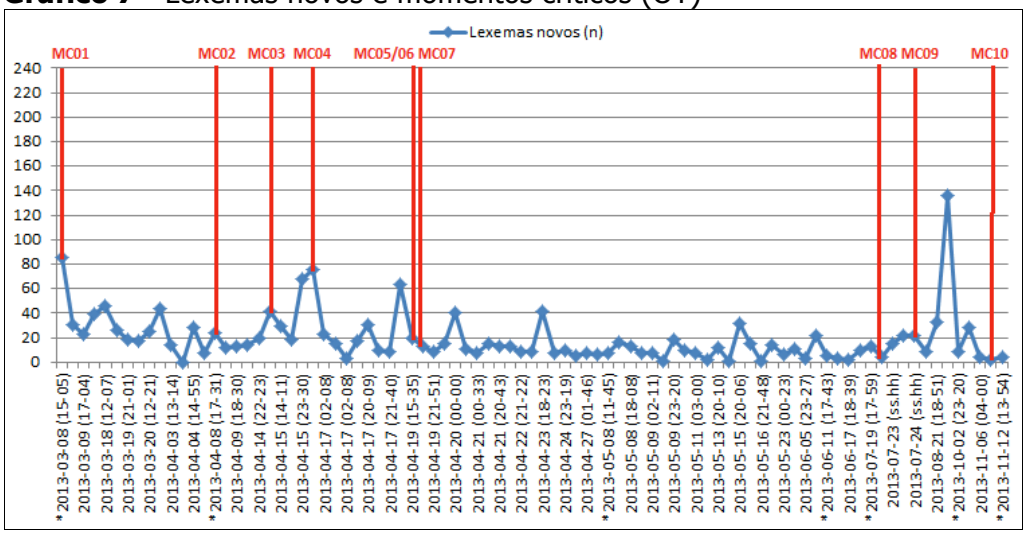
MC01: 08/03/13 Assassinato do primeiro repórter (R.N.)
MC02: 08/04/13 Protesto do Comitê R.N. pedindo apuração
MC03: 14/04/13 Assassinato do segundo repórter (W.)
MC04: 15/04/13 Constituição de força-tarefa para apuração das mortes
MC05: 15/04/13 Divulgação dos crimes no Vale do Aço investigados por R.N.
MC06: 19/04/13 Divulgação da informação de que houve policiais envolvidos nas mortes dos dois jornalistas
MC07: 19/04/13 Prisão de 2 primeiros policiais em uma série de 16 indiciados
MC08: 23/07/13 Divulgação do nome dos executores de R.N. e W.
MC09: 24/07/13 Divulgação dos executores de 7 crimes investigados por R.N.
MC10: 06/11/13 Divulgação dos executores de mais 2 crimes investigados por R.N.

Fonte: O autor.

Em síntese, o encadeamento dos eventos é o seguinte: o repórter R.N. foi assassinado por estar investigando crimes cometidos por policiais; o repórter W. teria tornado público que saberia quem teria matado R.N. e, por isso, também foi morto; as investigações identificaram os dois assassinos de R.N. e W. bem como os executores dos crimes investigados por R.N.

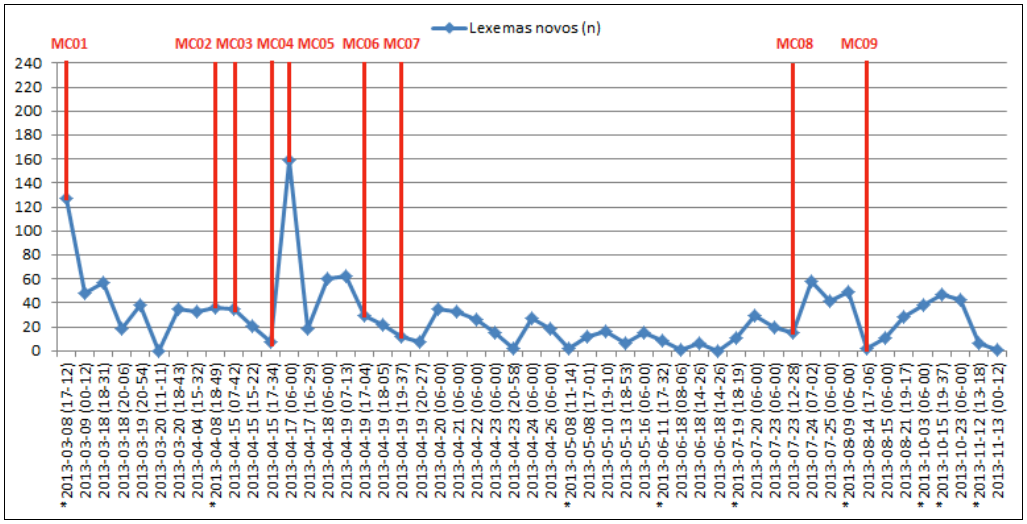
Vejam-se a seguir os gráficos que representam o número de lexemas novos e os momentos críticos dos eventos:

Gráfico 7 - Lexemas novos e momentos críticos (OT)



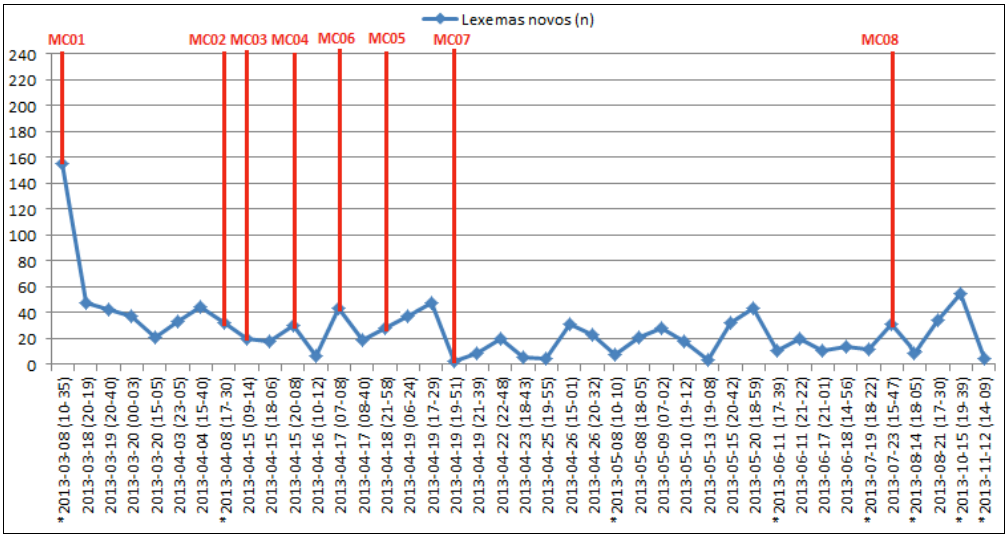
Fonte: O autor.

Gráfico 8 - Lexemas novos e momentos críticos (EM)



Fonte: O autor.

Gráfico 9 - Lexemas novos e momentos críticos (HD)



Fonte: O autor.

Antes de comentar os gráficos, convém dar algumas explicações:

(a) Em EM, M08 é genérico (fala-se no indiciamento de 16 pessoas sem individualizá-las); e

(b) Em HD, MC05 é genérico (fala-se em 14 homicídios na investigação, se individualizar cada um) e aparece apenas após MC06; não há MC09 e MC10.

Na análise dos gráficos, pode-se saltar inicialmente MC01, porque é o pico de informação nova nos três periódicos pelo simples fato de ser o início dos eventos que se tornarão em seguida tema das notícias.

Curiosamente, os momentos críticos em relação aos eventos não coincidem sistematicamente com os picos de lexemas novos (informação nova). Excluindo-se o MC01, tem-se em OT 4 coincidências em 10 (MC02, MC03, MC04 e MC09), em EM 3 coincidências (MC02, MC03 e MC05) e em HD também 3 coincidências (MC04, MC06 e MC08).

É interessante verificar que não existe, portanto, uma relação nítida entre informação nova e informação relevante. Isso demonstra que, ainda que um algoritmo baseado na análise do vocabulário seja capaz de quantificar informação nova, precisa-se de um procedimento auxiliar para identificar qual informação nova é relevante.

Ainda que a abordagem fundamentalmente quantitativa tenha falhado na identificação de informação relevante, ela pode sim fornecer subsídios para uma abordagem mais complexa no tratamento dos dados. Uma estratégia, a ser testada futuramente, seria a de identificar a informação nova (algo que os procedimentos aqui realizados permitem efetivamente) e extrair de cada texto a fração mínima, mas ainda assim interpretável, que contém essa informação nova: essa fração mínima poderia ser a frase em que está inserido o lexema novo ou ainda o parágrafo respectivo, limites facilmente identificáveis de forma automática (ponto final, ponto de interrogação, etc. para limite de frase; marca de parágrafo e/ou adentramento para parágrafo). Um tal algoritmo seria interessante para se excluir a informação velha considerando uma dada progressão de publicação de textos nos periódicos.

A técnica de quantificação de informação nova por via lexical poderá no futuro apresentar também contribuições para análises linguísticas funcionais, sobretudo para aquelas em que se dá ênfase à questão do fluxo de informação:

> [O] diferente estatuto informacional das diversas porções do texto corresponde a diferentes modos de codificação e de emissão [...], bem como a modos particulares de organização linear. O fluxo de informação determina a ordenação linear dos sintagmas nominais na frase, que se faz na sequência que o falante considera adequada para obter a atenção do ouvinte, mas alteração es da ordem pode atuar no sentido de controlar o fluxo da atenção. (NEVES, 1997, p. 35)

De modo geral, as análises funcionais tomam como parâmetro para identificação de informação velha os limites do texto (oral ou escrito) que está sendo investigado, mas com certeza o volume de informação conhecida de forma efetiva por cada interlocutor deve ser maior que o expresso formalmente naquele momento de interação, aspecto complicador para a realização de uma análise baseada no fluxo de informação.

A informática permite ir mais além dos limites da informação presente nas unidades textuais de um momento específico de interação, rastreando o conjunto de informações dadas sobre um tema. Assim, por exemplo, é de se esperar que a estruturação dos textos, sobre o tema aqui investigado,

publicados nos periódicos considerados tenha sido determinada pelo fluxo de informação, ou seja, a forma de cada texto deve ter sido determinada em função da informação já divulgada até então e, portanto, da informação velha. Essa expectativa é uma decorrência da adoção de uma visão funcionalista de linguagem e tem-se agora um instrumental técnico para ser realizada.

Considerações Finais

No presente estudo, realizou-se a análise do vocabulário de um conjunto de 179 textos, sobre um dado tema, publicados nos três periódicos de maior circulação, em Minas Gerais, ao longo de nove meses (março a dezembro de 2013). Essa análise foi realizada quantitativamente, levando em conta os seguintes critérios por periódico: número de textos (total e por mês), número de ocorrências de lexias (total e por mês), número de lexias diferentes (total e por mês), número de lexemas diferentes (total e por mês), número de lexemas novos por texto, número total acumulado de lexemas e porcentagem de lexemas novos por texto. Do ponto de vista qualitativo, fez-se uma comparação entre uma seleção de dez fatos relevantes, no curso dos eventos relativos ao tema considerado, e as medidas quantitativas acima listadas.

Como resultados, obteve-se que:

(a) os periódicos não se comportaram da mesma maneira frente à publicação de textos sobre o tema escolhido (OT publicou quase o dobro de EM e HD), o que permite inferir que a publicação reiterada de notícias sobre o tema seria uma estratégia de revelar um peso maior que um dado periódico atribui ao tema;

(b) dentre as estratégias de quantificação de informação nova através do vocabulário dos textos, a mais eficiente para diferenciar os textos é o número de lexemas diferentes;

(c) a quantidade de informação nova em cada novo texto, sobre o tema publicado em cada periódico, é relativamente baixa (gira em torno de 13% a 15%), fato que parece decorrer de dois aspectos: necessidade de

contextualizar a informação nova por meio de informação velha, por um lado, e limitação-padrão na extensão de cada texto, por outro;

(d) os periódicos ainda parecem vinculados a uma prática de organização da informação da época de mídia apenas impressa (com grande quantidade de informação velha por texto), apesar de já circularem em ambiente digital (que permitiria apresentar apenas a informação nova, ficando a velha disponível através de links), fato que talvez derive de uma realidade em que o redator de notícia para a mídia impressa seja o mesmo para a digital, não considerando ser necessária nenhuma adaptação, seja por não compreender as possibilidades de mídia digital seja por não considerar ser necessário dar-se a esse trabalho de diferenciação dos textos por tipo de mídia;

(e) embora a quantificação de informação nova através de contagem de lexemas novos tenha sido interessante para diferenciar a quantidade de informação por texto, não se mostrou suficiente para a identificação de informação relevante.

Em síntese, a quantificação de informação nova através de contagem de lexemas novo é relevante, mas precisa ser associado a um procedimento mais específico para ser capaz de considerar aspectos mais qualitativos como a relevância da informação nova.

Referências

BERG, Márcia Barreto. *O comportamento semântico-lexical das preposições do português do Brasil*. 2005. 161 f. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

BIDERMAN, Maria Tereza. A face quantitativa da linguagem: um dicionário de frequências do português. *Alfa*, São Paulo, v. 42, p. 157-181, 1998.

BIDERMAN, Maria Tereza. *Teoria linguística*: teoria lexical e linguística computacional. 2. ed. São Paulo: Martins Fontes, 2001.

CAMBRAIA, César Nardelli. Da lexicologia social a uma lexicologia sócio-histórica: caminhos possíveis. *Revista de Estudos de Linguagem*, Belo Horizonte, v. 21, n. 1,

p. 157-188, 2013. Disponível em: <http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/5096/4553>. Acesso em: 5 jun. 2016.

HOUAISS, Antônio et al. *Dicionário eletrônico Houaiss da língua portuguesa*. Rio de Janeiro: Objetiva, 2001.

MARCUSCHI, Luiz Antônio. O léxico: lista, rede ou cognição social? In: NEGRI, Lígia et al. (Org.). *Sentido e significação*: em torno da obra de Rodolfo Ilari. São Paulo: Contexto, 2004. p. 263-284.

MATORÉ, Georges. *La méthode en lexicologie*: domaine français. Paris: Didier, 1973.

NEVES, Maria Helena de Moura. *A gramática funcional*. São Paulo: Martins Fontes, 1997.