



Calidoscópico

E-ISSN: 2177-6202

calidoscopio@unisinis.br

Universidade do Vale do Rio dos Sinos  
Brasil

Silva Souza, Marco Aurélio; de Freitas, Maria Cláudia; Dias Pereira, Maria das Graças  
O léxico do trânsito do Rio de Janeiro: estudo da anotação de um corpus de reportagens  
aéreas do rádio

Calidoscópico, vol. 13, núm. 3, septiembre-diciembre, 2015, pp. 416-423  
Universidade do Vale do Rio dos Sinos

Disponível em: <http://www.redalyc.org/articulo.oa?id=571561401002>

- Como citar este artigo
- Número completo
- Mais artigos
- Home da revista no Redalyc

redalyc.org

Sistema de Informação Científica

Rede de Revistas Científicas da América Latina, Caribe, Espanha e Portugal

Projeto acadêmico sem fins lucrativos desenvolvido no âmbito da iniciativa Acesso Aberto

Marco Aurélio Silva Souza  
marcoarelio.professor@yahoo.com.br

Maria Cláudia de Freitas  
claudiafreitas@puc-rio.br

Maria das Graças Dias Pereira  
mgdpereira@terra.com.br

## O léxico do trânsito do Rio de Janeiro: estudo da anotação de um *corpus* de reportagens aéreas do rádio

### The lexicon of Rio de Janeiro traffic: Study of annotation of aerial reports corpus from the radio

**RESUMO** – Este artigo apresenta um estudo sobre o léxico do domínio “trânsito”, realizado por meio da anotação manual de um *corpus* de transcrições de reportagens aéreas sobre o trânsito, transmitidas em rádios da cidade do Rio de Janeiro. O objetivo da anotação é observar, de maneira sistemática, palavras e expressões usadas pelos repórteres aéreos ao se referir às diferentes situações do trânsito. O *corpus* é composto pela transcrição de 17 reportagens de quatro repórteres aéreos em seis rádios FM do Rio de Janeiro. As transcrições foram manualmente anotadas quanto às categorias e subcategorias do domínio “trânsito”, desenvolvidas ao longo da pesquisa, utilizando uma ferramenta de anotação, o Etiquet(H)AREM, especialmente adaptada para esse fim. Os resultados não só reforçam o imenso potencial da anotação como forma de estudo e de descrição de porções da língua, como também sistematizam um léxico de um gênero (reportagem aérea), ainda não estudado, com categorias do domínio “trânsito” e com informações que possibilitam a projeção futura das condições do tráfego em seus aspectos espacial, temporal e situacional.

**Palavras-chave:** *corpus*, anotação lexical, repórter aéreo, trânsito, rádio.

**ABSTRACT** – This article presents a study on the lexicon of the domain “traffic”, held by manual annotation of a corpus with transcriptions of aerial reports about the traffic, transmitted on radio stations of the city of Rio de Janeiro. The aim of the annotation is to observe, systematically, words and expressions used by aerial reporters referring to different situations on traffic. The corpus consists of the transcription of 17 reports from four aerial reporters in six FM radio stations in Rio de Janeiro. The transcriptions were manually annotated for assigning categories and subcategories of the domain “traffic”, developed during the research, using an annotation tool (Etiquet(H)AREM), especially adapted for this purpose. The results not only reinforce the immense potential of the annotation as a means of study and description of language portions as well as systematize a lexicon of a genre (aerial reports), not yet studied, with categories of the domain “traffic” and with information that enables the future projection of traffic conditions in their spatial, temporal and situational aspects.

**Keywords:** corpus, lexical annotation, aerial reporter, traffic, radio.

### Introdução

Apresentamos, neste artigo, um estudo sobre o léxico do trânsito, realizado por meio da anotação de um *corpus* de transcrições de reportagens aéreas transmitidas em rádios da cidade do Rio de Janeiro. Além de construir um recurso para o processamento automático de textos, o processo de anotação aqui conduzido teve como principal objetivo investigar detalhadamente os termos e expressões utilizados pelos repórteres aéreos para descrever a situação do trânsito e, simultaneamente, possibilitar um levantamento do

léxico deste domínio. Para tanto, anterior à anotação, precedeu-se ao desenho do esquema de anotação, que descreve as classes e subclasses relevantes para o estudo em questão.

Nesse contexto, a anotação permitiu a observação e a classificação das palavras utilizadas pelos repórteres aéreos para descrever o trânsito em diferentes situações. O *corpus* anotado e o léxico dele derivado podem, ainda, auxiliar sistemas ou ferramentas que lidem com o domínio “trânsito”, fornecendo dados importantes para a análise da situação do tráfego com informações para sistemas de previsão de fluxos na cidade.

Os congestionamentos são fenômenos urbanos que interferem diretamente em diversos setores da sociedade e na qualidade de vida dos cidadãos, e o trânsito em uma grande metrópole como o Rio de Janeiro recebe a atenção de diversos setores governamentais.

Para auxiliar os motoristas e cidadãos em seus deslocamentos em uma grande cidade, o serviço de reportagens aéreas é transmitido por emissoras de rádio e televisão de diversas cidades do mundo. Jornalistas e pilotos sobrevoam a cidade em helicópteros e observam diversos eventos urbanos. O foco de seus relatos, no entanto, é no trânsito de veículos pelas ruas, avenidas e estradas da Região Metropolitana. Os repórteres aéreos transmitem *flashes*, nos horários de *rush*, informando sobre congestionamentos nas vias, fornecendo aos motoristas opções de percursos em que o trânsito apresenta melhores condições. O objetivo do serviço é proporcionar ao motorista menos tempo de deslocamento, especialmente entre a casa e o trabalho e vice-versa. Os repórteres aéreos também transmitem informações que sejam relevantes para o fluxo do trânsito, como acidentes, manutenções nas vias, condições do tempo e de outros meios de transporte públicos, além de outras notícias de interesse jornalístico.

Uma vez que o trânsito tem se tornado um problema considerável no cotidiano de uma grande cidade, estudos dos campos lexicais de textos desta área podem contribuir para um melhor entendimento do que ocorre no fluxo do tráfego. Em uma perspectiva da Lexicologia Social (Cambraia, 2013), o fator social é primordial na organização do vocabulário. Nesse sentido, e considerando a relevância do trânsito para os moradores dos grandes centros urbanos, observar o léxico do trânsito – e a sua estruturação – é observar, também, uma parcela de nossa vida urbana.

Para conduzir o presente estudo, escolhemos a estratégia de anotação de *corpus* (Leech, 1997, por exemplo). Tradicionalmente vinculada à área da Linguística Computacional, a anotação, como advogam Sampson (2001), Archer (2012), Santos *et al.* (2015) e Freitas (2015), é, também, uma forma de estudo. Para o processo de anotação propriamente, etiquetagem, utilizamos a ferramenta Etiquet(H)AREM (Carvalho e Oliveira, 2008), que foi especialmente adaptada para dar conta do domínio “trânsito”.

Para este estudo, criamos um pequeno *corpus* e pudemos perceber e identificar, na língua em funcionamento, a constituição de um léxico convencional entre repórteres aéreos do Rio de Janeiro de diferentes rádios, demonstrando a utilização de um campo lexical bastante delimitado na transmissão dessas reportagens.

Os resultados não só reforçam o imenso potencial da anotação como forma de estudo e de descrição de porções da língua como também sistematizam um léxico de um gênero (reportagem aérea), ainda não estudado, com categorias do domínio “trânsito” e com informações que

possibilitam a projeção futura das condições do tráfego em seus aspectos espacial (localização), temporal (dia e hora do evento) e situacional (avaliação do trânsito).

Nesta seção, apresentamos uma breve descrição da pesquisa. Na segunda seção, descrevemos alguns estudos sobre a definição e tratamento de *corpus* e processos de anotação. Na terceira seção, apresentamos as características do *corpus* e as classes de anotação. Na quarta seção, descrevemos as categorias e apresentamos exemplos de aplicação das etiquetas.

## Corpus e anotação

A utilização de *corpus* – coleção de textos – como fonte de estudos (não necessariamente linguísticos) não é nova: data do século XIII e remete a estudos bíblicos e, posteriormente, literários (McCarthy e O’Keeffe, 2012 in Freitas, 2015). Mais recentemente, o termo *corpus* tem sido usado para se referir a qualquer coleção robusta de textos autênticos de linguagem natural e em formato eletrônico. Consideramos aqui a definição abrangente de Santos (2008), para quem um *corpus* é “uma coleção classificada de objetos linguísticos para uso em Processamento de Linguagem Natural/Linguística Computacional/Linguística”.

Um *corpus*, portanto, é, também, uma maneira de estudar (e explorar) a língua, ainda que, considerando os avanços e os interesses sobre a exploração de grandes coleções de documentos, a restrição ao âmbito estritamente linguístico e a *corpora* extensos esteja cada vez mais diluída (veja-se, por exemplo, Leite e Freitas, 2015 e Williford e Henry, 2012, para usos não “estritamente” linguísticos de *corpus*).

Já a anotação de *corpus* é a atribuição de rótulos, etiquetas ou metadados a segmentos de texto, que podem corresponder a palavras, locuções ou porções maiores, como parágrafos ou mesmo textos inteiros. Trata-se de uma prática que visa à inclusão de informação linguística interpretativa a um *corpus* eletrônico, de forma a enriquecê-lo, enquanto fonte de informação linguística para pesquisa e desenvolvimento futuros (Leech, 1997, 2005). A anotação também pode se referir ao produto final deste processo, com as etiquetas inseridas como representação de material linguístico. Em resumo, e de maneira simplificada, a anotação consiste tão somente em uma tarefa de classificação.

Outra dimensão da anotação de *corpus*, diretamente associada ao presente estudo, diz respeito à anotação como forma de estudo e pesquisa empírica (Sampson, 2001; Archer, 2012; Freitas, 2015; Santos *et al.*, 2015). Neste caso, a anotação materializa um estudo empírico sobre porções da língua. Na anotação manual ou semiautomática, quando o pesquisador é responsável não apenas pela anotação, mas também pelo desenho do esquema de anotação – as classes (ou etiquetas) usadas para anotar –,

ele está diante de um valioso processo empírico de investigação (Freitas, 2015; Archer, 2012).

Ao participar do desenvolvimento de um projeto de anotação, o pesquisador é obrigado a se apropriar do texto-conteúdo que será anotado e das classes que compõem o esquema de anotação, o que se configura como um teste empírico desenhado de forma clássica, em que se criam hipóteses (categorias provisórias), que são verificadas (aplicadas no *corpus*; o processo de anotação propriamente). Durante este processo, as hipóteses-categorias podem ser confirmadas (a anotação ocorre como previsto) ou, por outro lado, os dados podem levar à reformulação das categorias iniciais. As novas categorias (hipóteses) são verificadas novamente no *corpus*, e o processo recomeça (Freitas, 2015). Ao longo do presente trabalho, a anotação materializa e possibilita um estudo e levantamento sobre os termos usados no domínio “trânsito”.

Em consonância com o objetivo da presente pesquisa, Archer (2012) afirma que o processo de anotação manual “tende a não ser exaustivo”, uma vez que, após a observação do pesquisador, “somente as palavras, sentenças e enunciados que são diretamente relevantes para a pesquisa são anotadas”. Conforme a autora, o nível de detalhamento de um esquema de anotação depende do tipo de texto e do objetivo da pesquisa.

### Aspectos metodológicos

Nesta seção, descrevemos as características do *corpus* e as classes de anotação.

Para estudar a linguagem dos repórteres aéreos no domínio “trânsito”, partimos de um *corpus* composto por 17 reportagens aéreas gravadas e transcritas, de quatro diferentes repórteres aéreos, transmitidas entre maio de 2011 e março de 2012, em seis rádios FM do Rio de Janeiro, totalizando 2.891 mil palavras. As gravações em áudio nas rádios foram realizadas no computador, a partir das páginas das emissoras na Internet, e através de sintonização FM das rádios em *smartphone*. Os arquivos foram armazenados em computador nos formatos MP3 e AMR.

O número de reportagens, de rádios e de repórteres é representativo do universo pesquisado, uma vez que cobre todas as rádios e repórteres que realizavam o serviço na época em que foram gravados e pelo fato de não termos verificado diferenças entre as reportagens transcritas e as atuais. Além disso, “quando a anotação é feita privilegiando o estudo linguístico, a exigência da quantidade perde a relevância” (Santos *et al.*, 2015).

As transcrições são analisadas e anotadas manualmente, em um processo de atribuição contextualizada de etiquetas criadas especificamente para as categorias do domínio trânsito. Como não se trata de um grande projeto de anotação de *corpus*, mas da utilização do processo de anotação como fonte de estudo, a inexistência da aferição

da concordância entre anotadores não é um problema, como advogado em Santos *et al.* (2015).

Diferentemente da “presente proliferação e facilidade de obter todos os tipos de gêneros de texto” (Santos, 2008, p. 45) na Internet, não foram encontrados *corpora* deste gênero disponíveis para consulta, e fez-se necessária a criação deste pequeno *corpus*, com informação linguística original e que atende aos objetivos da pesquisa.

Em relação à tipologia do texto analisado, este gênero (reportagem aérea), transmitido por rádio ou televisão, é produzido exclusivamente em linguagem oral transcrita, é contemporâneo, de amostragem, com conteúdo especializado, de língua nativa, com a finalidade de ser objeto de estudo linguístico e social e para teste de ferramentas futuras de análise (Sardinha, 2000, p. 340).

Após a compilação do *corpus*, procedeu-se à definição das categorias de anotação. As categorias [C:01TRAFEGO], [C:02LOCAL] e [C:03AVALIACAO] foram determinadas de acordo com o léxico utilizado pelos repórteres aéreos para descrever o tráfego, o local do evento e a avaliação da situação. Os critérios de anotação foram determinados pela orientação específica ao tema do objeto de estudo (o trânsito). Apenas as palavras/sentenças relacionadas ao tema foram anotadas. As três categorias, após a análise do texto, foram divididas em treze etiquetas (Quadro 1).

Essas categorias-hipóteses se mostraram adequadas e produtivas para classificar as palavras pertencentes ao domínio aqui analisado. Evidentemente, nem todas as palavras se encaixam perfeitamente em um determinado campo semântico, e algumas podem sinalizar mais de um conceito dentro de um mesmo texto (Santos, 2007, p. 45). Contextos específicos, linguísticos e sociais, como o do

**Quadro 1.** Etiquetas para anotação lexical do *corpus*.  
**Chart 1.** Tags for lexical annotation of the corpus.

Categorias (C), Tipos (T), Subtipo (S)
C:01TRAFEGO
T:01aTRAFEGOALVO
T:01bTRAFEGOSUBALVO
C:02LOCAL
T:02aLOCALCOMPLETO
T:02bLOCALPARTE
S:02baLOCALSUBPARTE
C:03AVALIACAO
T:03aAV--
T:03aAV-
T:03bAV++
T:03bAV+
T:03cAVN

trânsito neste estudo, determinam diferentes conceitos para as palavras. Estes conceitos definem o “modelo de organização e funcionamento do léxico, [que] não pode prescindir da articulação de fatores intralinguísticos e extralinguísticos” (Camraia, 2013, p. 167-168).

Nesse sentido, a anotação conduzida levou em conta, sempre, o contexto. Palavras ou expressões que pertenceriam ao domínio “trânsito” que, no contexto, indicam a situação, a localização e a avaliação do tráfego pelo repórter aéreo. Por exemplo, as expressões “muita calma nessa hora” e “show de bola” podem ser utilizadas em diferentes contextos para expressar opinião. No contexto do trânsito, são utilizadas para indicar, respectivamente, uma avaliação muito negativa do trânsito e uma avaliação muito positiva do trânsito. Desta forma, as expressões mencionadas, ao serem analisadas dentro deste contexto específico, passam a compor o léxico do trânsito.

As treze etiquetas utilizadas para a anotação deste *corpus* compõem as três categorias, nove tipos e o subtipo (Quadro 1), de modo a agrupar palavras e expressões relacionadas e permitir a anotação do maior número possível de itens.

Para realizar a anotação, adaptamos o Etiquet (H)AREM, uma ferramenta de auxílio à anotação de *corpora*, desenvolvida para a anotação semântica de entidades mencionadas (EMs) e de relações entre EMs, no âmbito do HAREM (Figura 1). O HAREM, por sua vez, “é uma avaliação conjunta na área do reconhecimento de entidades mencionadas em português. De forma simplificada, trata-se de uma iniciativa que pretende avaliar o sucesso na identificação e consequente classificação automática dos nomes próprios na língua portuguesa” (Carvalho e Oliveira, 2008, p. 340).

## Anotação das reportagens aéreas sobre o trânsito

Nesta seção, descrevemos as categorias e apresentamos exemplos, retirados do *corpus*, de aplicação das etiquetas.

### Categoria [C:01TRAFEGO]

A categoria [C:01TRAFEGO] se refere aos itens lexicais que apresentam o início do trecho de informação sobre o trânsito. Deve ser utilizada quando o anotador inferir que um item lexical indica que o repórter aéreo está se referindo especificamente a alguma informação acerca do deslocamento. Em uma frase como [O trânsito na saída do centro do Rio tá bem complicado.], anotamos com essa categoria a palavra [trânsito].

A categoria é dividida em dois tipos: [T:01aTRAFEGOALVO] e [T:01bTRAFEGOSUBALVO]. O tipo [T:01aTRAFEGOALVO] deve ser utilizado para anotação quando o item lexical se referir diretamente ao trânsito, como no exemplo [A movimentação pela Niemeyer ainda é um pouco melhor.]. Neste caso, anotamos a palavra [movimentação] por se referir ao trânsito e por poder ser substituída por outras palavras do mesmo campo lexical (Tabela 1). O tipo [T:01bTRAFEGOSUBALVO] deve ser utilizado para anotação quando o item lexical se referir indiretamente ao trânsito, como na frase [Por esse trajeto a galera vai perder um tempinho.]. Anotamos a palavra [trajeto] como referência indireta ao trânsito, não se configurando como o local da avaliação ou à avaliação, mas como uma subparte do trânsito, uma meronímia que os tipos possuem em relação às suas categorias.

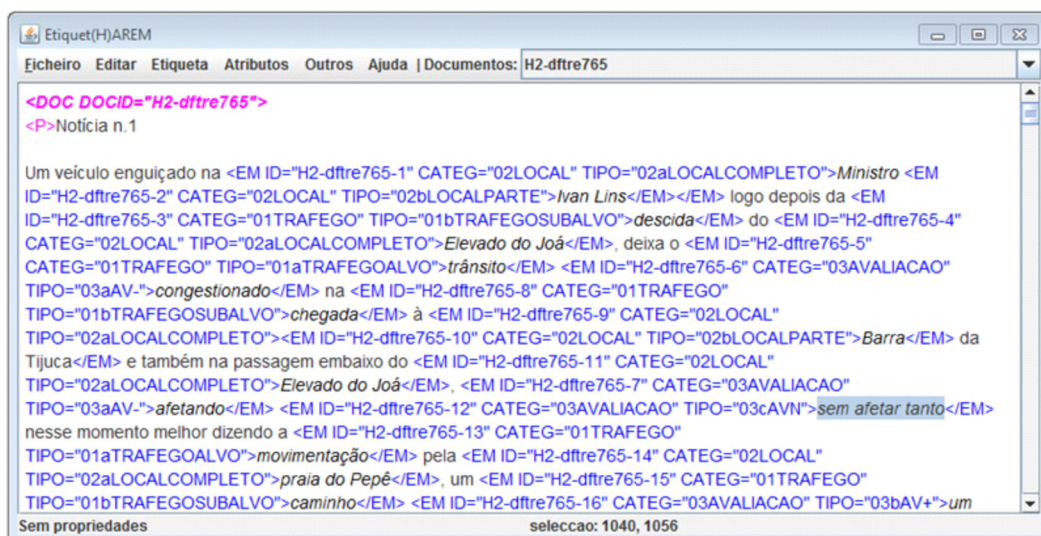


Figura 1. Exemplo da anotação das categorias no Etiquet(H)AREM.

Figure 1. Example of annotation of categories in Etiquet(H)AREM.



### **Categoria [C:02LOCAL]**

Os itens lexicais que se referem ao local físico em que o repórter aéreo está descrevendo a situação devem ser anotados com a categoria [C:02LOCAL], que é dividida em dois tipos, [T:02aLOCALCOMPLETO] e [T:02bLOCALPARTE], e um subtipo [S:02baLOCALSUBPARTE] (Tabela 2).

A anotação da categoria compreende, portanto, os nomes dos logradouros (nomes de ruas, avenidas, estradas, praças, logradouros e seus trechos, bairros, regiões e cidades), como no exemplo: [Quem tá trafegando pela Linha Vermelha ou pela Avenida Brasil em direção à Baixada, em direção à Zona Oeste, enfrenta alguns probleminhas]. Neste caso, anotamos [Linha Vermelha], [Avenida Brasil], [Baixada] e [Zona Oeste].

Entretanto, a referência que o repórter aéreo faz ao nome de um logradouro que é composto por mais de uma palavra pode ser realizada com seu nome completo ou com parte do nome. Ou seja, o nome composto de um logradouro pode ser desmembrado. No exemplo [Quem tá trafegando pela Linha Vermelha], o nome da via [Linha Vermelha], em alguns casos, é mencionado no *corpus* somente como [Vermelha], como no exemplo [Nem adianta cair pros lados da Vermelha].

Nos casos de locais que apresentam nomes compostos, estes foram anotados com a categoria [C:02LOCAL] e o tipo [T:02aLOCALCOMPLETO] para o nome composto, e com o tipo [T:02bLOCALPARTE] para o nome simples que representa o nome composto do logradouro. No caso de [Avenida Brasil], o nome simples [Brasil] representa o nome composto, como no exemplo [Trânsito ruim agora nos acessos tanto pela Brasil como pela Linha Vermelha]. Outro exemplo seria [Pela Brasil é uma melhor opção], onde [Brasil] se refere à Avenida Brasil e deve ser anotado com o tipo [T:02bLOCALPARTE].

Portanto, nos exemplos acima, além de marcarmos os logradouros que possuem nomes compostos com a categoria [C:02LOCAL], anotamos também com os tipos [T:02aLOCALCOMPLETO], por apresentarem nomes compostos, e [T:02bLOCALPARTE], para os nomes simples que representam o nome composto [Vermelha] e [Brasil]. O local [Baixada] foi anotado somente com o

tipo [T:02bLOCALPARTE], pois apresenta somente parte do nome composto [Baixada Fluminense]. O objetivo da anotação múltipla de logradouros com nomes compostos é auxiliar ferramentas automáticas a identificarem os locais mesmo que suas denominações estejam incompletas.

Além de um logradouro ter seu nome parcialmente mencionado, como [Ponte] por [Ponte Rio-Niterói], trechos ou características do logradouro podem ser mencionados como referência ao logradouro. No exemplo [Na Ponte há dificuldades em praticamente todo o deslocamento mas principalmente na ( ) do Vão central, até a chegada à Praça do pedágio], anotamos os itens [Vão central] e [Praça do pedágio] com a categoria [C:02LOCAL], com o tipo [T:02bLOCALPARTE] e com o subtipo [S:02baLOCALSUBPARTE], pois são reconhecidos pelos motoristas como partes do logradouro Ponte Rio-Niterói e possuem referência geográfica definida e grande relevância na descrição e avaliação da situação do trânsito naquele local.

### **Categoria [C:03AVALIACAO]**

Após as referências ao local, o repórter aéreo realiza uma avaliação da situação do trânsito. Em nosso *corpus*, anotamos estas avaliações com a categoria [C:03AVALIACAO]. No exemplo [Trânsito livre no Aterro do Flamengo], anotamos o item [livre], pois apresenta uma avaliação do modo como o trânsito está fluindo no local mencionado [Aterro do Flamengo].

Esta categoria possui uma característica fundamental na anotação, pois é essencial para o ouvinte saber exatamente as condições do tráfego nos locais pelos quais deseja transitar. Dividimos em 5 tipos, que representam gradativamente a avaliação que o repórter aéreo faz da situação – de trânsito muito ruim [T:03aAV--] a trânsito muito bom [T:03bAV++] (Tabela 3).

O tipo [T:03aAV--] representa uma situação muito negativa do trânsito, como em [Péssimas condições, em todas estas vias.] em que marcamos o item [Péssimas] como significativo de uma situação muito ruim.

O tipo [T:03aAV-] representa uma situação relativamente negativa e apresentou maior número de ocorrências nesta categoria (Tabela 4). No exemplo [O movimento é intenso nesse horário, com retenções na Autoestrada.],

**Tabela 1.** Campo lexical da categoria [C:01TRAFEGO].  
**Table 1.** Lexical field of the category [C:01TRAFEGO].

<b>Categoria</b>	<b>Níveis de anotação</b>	<b>Campo lexical</b>
C:01TRAFEGO	T:01aTRAFEGOALVO	condições, desloca, deslocamento, movimento, movimentação, situação, trânsito, tráfego, trafega, trafegando
	T:01bTRAFEGOSUBALVO	acesso, acessos, altura, caminho, caminhos, chegada, destino, direção, descida, extensão, percurso, subida, trajeto

**Tabela 2.** Campo lexical da categoria [C:02LOCAL].**Table 2.** Lexical field of the category [C:02LOCAL].

<b>Categoria</b>	<b>Níveis de anotação</b>	<b>Campo lexical</b>
C:02LOCAL	T:02aLOCALCOMPLETO	Alameda São Boaventura, Avenida Airton Senna, Avenida das Américas, Avenida Brasil, Avenida Lúcio Costa, Avenida Niemeyer, Avenida Paulo de Frontin, Autoestrada Lagoa-Barra, Elevado da Perimetral, Elevado do Joá, Estrada Velha da Barra da Tijuca, Enseada de Botafogo, Largo do Humaitá, Linha Amarela, Linha Vermelha, Ponte Rio-Niterói, Túnel Rebouças, Barra da Tijuca, Botafogo, Copacabana, Niterói, Nikiti
	T:02bLOCALPARTE	Alameda, Américas, Brasil, Lúcio Costa, Niemeyer, Autoestrada, Lagoa-Barra, Perimetral, Amarela, Vermelha, Ponte, Rio-Niterói, Rebouças, Barra, Recreio, Centro, Centro do Rio, Rio, Maxwell, Jansen de Melo, Francisco Bicalho, Refinaria
	S:02baLOCALSUBPARTE	vão central, grande curva, praça do pedágio, pista, pistas, pista central, pista lateral, lateral, faixa, faixas, faixa de rolamento, retorno

**Tabela 3.** Campo lexical da categoria [C:03AVALIACAO].**Table 3.** Lexical field of the category [C:03AVALIACAO].

<b>Categoria</b>	<b>Níveis de anotação</b>	<b>Campo lexical</b>
C:03AVALIACAO	T:03aAV--	afetando bastante, bem complicado, bem complicada, bem ruim, lentidão, muito complicado, muito grande, muito mal, piora, perde muito tempo, perdendo muito tempo, péssimas condições
	T:03aAV-	acidente, afetando, atenção, complicado, complicada, complicações, congestionado, congestionamento, difícil, dificuldade, dificuldades, enguiçado, intenso, intensa, interdita, lento, mal, manifestação, muita calma nessa hora, não é boa, nada diferente, nada bom, não são boas, não é das melhores, paciência, parado, perde tempo, perder um tempinho, pequenas retenções, problema, problemas, probleminhas, retenção, retenções, ruim
	T:03bAV++	facilidade, melhora bastante, melhorou bastante, show de bola
	T:03bAV+	bom, boa, boas, boas condições, fluindo, liberou, livre, melhor, melhores, melhorou, melhora, não tá tão ruim, sem atrapalhar, menor, tá legal, um pouco melhor
	T:03cAVN	normal, razoável, sem afetar tanto

anotamos [intenso] e [retenções] como avaliações pouco negativas da situação do tráfego no local mencionado.

O tipo [T:03bAV+] representa uma situação parcialmente positiva. Em [Pela Brasil é uma melhor opção.], anotamos o item [melhor], que demonstra uma boa situação do tráfego.

O tipo [T:03bAV++] representa uma situação muito positiva do trânsito. No exemplo [A situação melhora bastante em direção à Barra da Tijuca.], anotamos a

expressão [melhora bastante], que significa uma situação de tráfego fluente.

A avaliação do repórter aéreo pode se referir também a uma situação sem avaliação positiva ou negativa explícita do trânsito, e o ouvinte deve inferir. Em [Na Avenida Maracanã o trânsito até que é razoável.], marcamos o item razoável com o tipo [T:03cAVN], deixando para o ouvinte a interpretação de uma situação nem boa nem ruim do tráfego.

Em relação às palavras anotadas, alguma informação pode ficar perdida ou ser marcada de modo inexato no processo de anotação, que é, sempre e inevitavelmente, um processo interpretativo. Por exemplo, o item [razoável] pode ser interpretado como indicativo de que há algum trânsito, mas que é suportável. Ademais, consideramos que “não há anotação perfeita do mesmo modo que não há interpretação perfeita” (Freitas, 2015).

Por fim, os itens lexicais podem corresponder a uma palavra ou a um conjunto de palavras. Por exemplo, uma avaliação pouco negativa [T:03aAV-] da situação do tráfego em determinado local pode ser expressa por uma palavra, [congestionado] ou [retenção], ou por mais de uma palavra, [meio lento] ou [perder um tempinho]. No entanto, [lento] ou [muito lento] significariam uma situação muito negativa [T:03aAV--].

A quantificação da categoria [C:03AVALIACAO] mostra que o número de avaliações negativas [T:03aAV-] é predominante, em contraste com o número de avaliações positivas [T:03aAV+] (Tabela 4). Esta verificação sinaliza a nossa relação negativa com o trânsito na grande maioria das vezes. Em outro aspecto, a situação do tráfego nos horários de pico pode ser avaliada pelo pequeno número de avaliações muito positivas [T:03aAV++].

## Conclusões

Neste trabalho, apresentamos um *corpus* criado para o estudo do domínio “trânsito”, descrevemos o seu processo de anotação e indicamos os resultados da

anotação-estudo. O *corpus* é composto de transcrições de reportagens aéreas transmitidas em rádios da cidade do Rio de Janeiro.

Os objetivos consistiram em detalhar os termos e expressões utilizados no gênero reportagem aérea e investigar o campo lexical do domínio “trânsito”. Contextualizada nas descrições realizadas em tempo real pelos repórteres aéreos, a anotação permitiu o estudo e a elaboração de um léxico específico do domínio “trânsito”, o que não apenas contribui para um retrato mais detalhado de uma das dimensões de nossa vida em sociedade, como também pode, em associação a sistemas dedicados ao trânsito, possibilitar a identificação de locais com melhores ou piores condições.

Ao longo do processo de anotação, verificamos que poucas palavras não foram anotadas (Figura 1). Isto significa que o léxico utilizado pelo repórter aéreo, quando analisado dentro do contexto do trânsito, mostra-se semanticamente contextualizado à atividade. Indica, ainda, que a fala neste contexto é altamente densa e informativa, sobrando pouco espaço para outros temas.

Foram anotados 966 itens lexicais (Tabela 4). Os itens anotados mostraram-se recorrentes em todas as transcrições que compõem o *corpus*, demonstrando a estabilidade do fenômeno observado.

A escolha do gênero reportagem aérea se deu pela grande quantidade de informação associada ao gênero e pela possibilidade de, através da anotação, extrair dele essa informação, de forma a contribuir para o desenvolvimento de ferramentas que auxiliem na compreensão

**Tabela 4.** Número de ocorrências dos itens lexicais anotados.

**Table 4.** Number of occurrences of annotated lexical items.

<b>Categorias (C), Tipos (T) e Subtipo (S)</b>	<b>Ocorrências</b>	<b>Percentual</b>
<b>C:01TRAFEGO</b>	<b>198</b>	<b>100%</b>
T:01aTRAFEGOALVO	103	52,0%
T:01bTRAFEGOSUBALVO	95	48,0%
<b>C:02LOCAL</b>	<b>535</b>	<b>100%</b>
T:02aLOCALCOMPLETO	245	45,8%
T:02bLOCALPARTE	272	50,9%
S:02baLOCALSUBPARTE	18	3,3%
<b>C:03AVALIACAO</b>	<b>233</b>	<b>100%</b>
T:03aAV--	19	8,1%
T:03aAV-	179	76,8%
T:03bAV++	4	1,7%
T:03bAV+	27	11,5%
T:03cAVN	4	1,7%
<b>Total</b>	<b>966</b>	



do funcionamento da questão social relacionada aos congestionamentos.

O presente estudo, um piloto, pode servir como base para outras pesquisas no âmbito da anotação de termos relacionados ao trânsito. Adicionalmente, o *corpus* compilado permite o estudo de uma das linguagens produzidas no ambiente radiofônico. Não temos notícia de outros *corpora*, em língua portuguesa, que contemplem este gênero, e, assim, também esperamos contribuir com mais material para o estudo da nossa língua e de nossas práticas.

## Referências

- ARCHER, D. 2012. Corpus annotation: a welcome addition or an interpretation too far? In: J. TYRKKÖ; M. KIPPIÖ; T. NEVALAINEN; M. RISSANEN (eds.), *Outposts of Historical Corpus Linguistics: from the Helsinki corpus to a proliferation of resources*. Varieng - Studies in Variation, Contacts and Change in English, 10. Disponível em: <http://www.helsinki.fi/varieng/series/volumes/10/archer/>. Acesso em: 26/11/2015.
- CAMBRAIA, C.N. 2013. Da lexicologia social a uma lexicologia sócio-histórica: caminhos possíveis. *Revista de Estudos de Linguagem*, 21(1):157-188. <http://dx.doi.org/10.17851/2237-2083.21.1.157-188>
- CARVALHO, P.; OLIVEIRA, H.G. 2008. Manual do Etiket(H)AREM. In: C. MOTA; D. SANTOS (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: o segundo HAREM*. Linguatca. Apêndice F, p. 339-346. Disponível em: <http://www.linguatca.pt/LivroSegundoHAREM/>. Acesso em: 26/11/2015.
- FREITAS, C. 2015. Corpus, Linguística Computacional e as Humanidades Digitais. In: M. LEITE; C.T. GABRIEL (orgs.), *Linguagem, Discurso, Pesquisa e Educação*. Rio de Janeiro, DP et ali, p. 18-46.
- HAREM. 2008. HAREM: Reconhecimento de entidades mencionadas em português. Disponível em: <http://www.linguatca.pt/HAREM/>. Acesso em: 13/11/2015.
- LEECH, G. 1997. Introducing corpus annotation. In: R. GARSIDE; G. LEECH; A. MCENERY (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, p. 1-18.
- LEECH, G. 2005. Adding Linguistic Annotation. In: M. WYNNE (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford, Oxbow Books, p. 17-29. Disponível em: <http://ahds.ac.uk/linguistic-corpora/>. Acesso em: 09/11/2015.
- LEITE, M.; FREITAS, C. 2015. Pesquisa em Educação: perspectivas (qualitativas?) na exploração de grandes corpora. *OSLa-Oslo Studies in Language*, 7(1):139-152.
- SAMPSON, G. 2001. *Empirical Linguistics*. London, Continuum, 226 p.
- SANTOS, D. 2007. O modelo semântico usado no Primeiro HAREM. In: D. SANTOS; N. CARDOSO (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguatca, p. 43-57. Disponível em: <http://www.linguatca.pt/LivroHAREM/>. Acesso em: 26/11/2015.
- SANTOS, D. 2008. Corporizando algumas questões. In: S. TAGNIN; O. VALE (orgs.), *Avanços da Linguística de Corpus no Brasil*. São Paulo, Editora Humanitas/FFLCH/USP, p. 41-66.
- SANTOS, D.; MARQUES, R.; FREITAS, C.; MOTA, C.; SIMÕES, A. 2015. Comparando anotações linguísticas na Gramatca: filosofia, ferramentas e exemplos. *Revista Domínios de Linguagem*, 9(2):11-26. <http://dx.doi.org/10.14393/DL18-v9n2a2015-2>
- SARDINHA, T.B. 2000. Linguística de Corpus: histórico e problemática. *DELTA*, 16(2):323-367. Disponível em: [http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S0102-44502000000200005&lng=pt&nrm=iso&tlng=pt](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=pt&nrm=iso&tlng=pt). Acesso em: 13/11/2013. <http://dx.doi.org/10.1590/S0102-44502000000200005>
- WILLIFORD, C.; HENRY, C. 2012. *One Culture: Computationally Intensive Research in the Humanities and Social Sciences – A Report on the Experiences of First Respondents to the Digging into Data Challenge*. Washington, D.C., Council on Library and Information Resources, 45 p. Disponível em: <http://www.clir.org/pubs/reports/pub151>. Acesso em: 26/11/2015.

Submetido: 30/07/2015

Aceito: 27/11/2015

### Marco Aurélio Silva Souza

Pontifícia Universidade Católica do Rio de Janeiro  
Rua Marquês de São Vicente, 225  
22453-900, Rio de Janeiro, RJ, Brasil

### Maria Cláudia de Freitas

Pontifícia Universidade Católica do Rio de Janeiro  
Rua Marquês de São Vicente, 225  
22453-900, Rio de Janeiro, RJ, Brasil

### Maria das Graças Dias Pereira

Pontifícia Universidade Católica do Rio de Janeiro  
Rua Marquês de São Vicente, 225  
22453-900, Rio de Janeiro, RJ, Brasil