

Metodologia de constituição de *corpora* orais: reflexões no âmbito da Linguística Funcional Centrada no Uso

Methodology for the construction of oral *corpora*: reflections within Usage-Based Functional Linguistics

Patrícia Fabiane Amaral da Cunha Lacerda¹, Lauriê Ferreira Martins Dall'Orto²

Universidade Federal de Juiz de Fora, Brasil

RESUMO

Este artigo discute o problema metodológico da constituição de *corpora* orais a partir de transcrições de vídeos do YouTube, considerando sua adequação a estudos baseados em frequência na Linguística Funcional Centrada no Uso. A relevância do estudo reside na necessidade de maior rigor metodológico no tratamento de dados empíricos em pesquisas dessa natureza. Nesse sentido, objetiva-se propor diretrizes para a constituição de *corpora* controlados por número de palavras, contemplando procedimentos que assegurem maior sistematicidade e comparabilidade aos dados analisados. Metodologicamente, trata-se de um estudo teórico-metodológico que combina revisão crítica da literatura e proposição de protocolo sistemático, com integração de técnicas qualitativas e quantitativas. Como resultados, indicam-se parâmetros para maior confiabilidade analítica, evidenciando o papel central da constituição do *corpus* na interpretação de padrões de uso e processos de mudança linguística.

PALAVRAS-CHAVE:

Linguística Funcional Centrada no Uso. *Corpora* orais. Frequência de uso. Método misto. Dados do YouTube.

ABSTRACT

This article addresses the methodological problem of constructing oral *corpora* from transcriptions of YouTube videos, considering their suitability for frequency-based studies within Usage-Based Functional Linguistics. The relevance of the study lies in the need for greater rigor in the selection, organization, and control of empirical data in this approach. The aim is to propose guidelines for compiling *corpora* controlled by word count, articulating criteria for data selection, organization, and analysis. Methodologically, this is a theoretical-methodological study that combines a critical review of the literature with the proposal of a systematic protocol, integrating qualitative and quantitative techniques. The results indicate parameters for greater analytical reliability, highlighting the central role of *corpus* construction in interpreting usage patterns and processes of linguistic change.

KEYWORDS:

Recebido em: 30/04/2026

Aceito em: 14/08/2026

¹ E-mail: | ORCID: patricia.cunha@ufjf.br | <https://orcid.org/0000-0002-0970-224X>.

² E-mail: | ORCID: martins.laurie@ufjf.br | <https://orcid.org/0000-0001-8864-8626>.

Usage-Based Functional Linguistics. Spoken *corpora*. Frequency of use. Mixed methods. YouTube data

1. Introdução

Nas últimas décadas, a Linguística Funcional Centrada no Uso (Martelotta, 2011; Furtado da Cunha *et al.*, 2013; Rosário; Oliveira, 2016; Bispo; Silva, 2016) tem se consolidado como uma produtiva abordagem teórica para a investigação linguística, ao conceber a língua como uma rede de construções emergente a partir do uso. Nesse quadro, a análise linguística passa a depender diretamente da observação empírica de dados reais de interação comunicativa, o que confere aos *corpora* um papel central tanto na descrição quanto na explicação de fenômenos linguísticos. A frequência de ocorrência, nesse contexto, assume estatuto teórico relevante, sendo frequentemente mobilizada como evidência para processos de mudança.

Embora haja um reconhecimento amplo do papel da frequência e do uso na modelagem da gramática e do léxico, observa-se que a discussão metodológica sobre a constituição dos *corpora* ainda permanece, em muitos casos, em segundo plano. Em outras palavras, enquanto se tem avançado significativamente na elaboração de modelos teóricos e na aplicação de métodos de análise – especialmente a partir da articulação entre abordagens qualitativas e quantitativas –, pouca atenção tem sido dedicada à problematização dos critérios que orientam a seleção, a organização e o controle dos dados que compõem os *corpora*. Tal lacuna torna-se particularmente relevante quando se considera que a validade das generalizações extraídas das análises quantitativas depende diretamente da comparabilidade e da representatividade dos dados analisados.

No âmbito da Linguística, o *corpus* pode ser definido como um conjunto sistematicamente organizado de dados linguísticos, selecionados segundo critérios explícitos e representativos de determinado uso da língua, com o objetivo de possibilitar sua descrição e análise empírica (McEnery; Hardie, 2012). Essa definição implica compreender o *corpus* não como uma coleção aleatória de ocorrências, mas como uma construção metodologicamente orientada, cuja constituição envolve decisões teóricas e analíticas que impactam diretamente os resultados da pesquisa.

Sob a perspectiva da Linguística Funcional Centrada no Uso, o *corpus* assume um papel ainda mais central, uma vez que a língua é concebida como uma rede de construções emergente do uso efetivo. Nesse sentido, conforme argumentam Traugott e Trousdale (2013), os padrões

linguísticos são abstraídos a partir da recorrência de construtos em contextos reais de interação, sendo a frequência um fator fundamental para processos de mudança. Assim, o *corpus* constitui o espaço empírico em que tais padrões podem ser observados e analisados.

Além disso, como destaca Vitral (2006), a constituição do *corpus* deve obedecer a critérios de controle que garantam a comparabilidade entre os dados, especialmente no que se refere ao número de palavras, uma vez que a frequência só pode ser interpretada em relação ao universo textual analisado. Nesse sentido, o *corpus* não é apenas fonte de dados, mas também uma unidade de medida, uma vez que delimita o contexto quantitativo dentro do qual as ocorrências são interpretadas. De forma complementar, Cunha Lacerda (2016), ao propor a articulação entre análise qualitativa e quantitativa no âmbito do método misto, evidencia que o *corpus* deve ser constituído de modo a permitir tanto a descrição contextual dos usos quanto a mensuração de sua frequência e distribuição. Sua delimitação, entretanto, nem sempre pode ser estabelecida previamente de maneira rígida, sobretudo quando o pesquisador ainda não dispõe de parâmetros suficientes para determinar a quantidade de dados necessária à compreensão do fenômeno investigado. Nesses casos, o critério de saturação pode constituir um recurso metodológico relevante, na medida em que permite avaliar se a incorporação de novos dados continua a acrescentar informações analiticamente significativas. A constituição do *corpus*, portanto, mostra-se indissociável das estratégias analíticas adotadas e dos critérios empregados para definir sua extensão, atuando como elemento estruturante do processo investigativo. Nessa perspectiva, o *corpus* configura-se simultaneamente como fonte de dados, unidade de análise e instrumento metodológico, cuja organização sistemática constitui condição fundamental para a produção de análises empiricamente fundamentadas e teoricamente consistentes. Nesse sentido, este artigo tem como objeto de análise a metodologia de constituição de *corpora* orais, com foco específico na utilização de transcrições de vídeos do YouTube no âmbito da Linguística Funcional Centrada no Uso. Parte-se do entendimento de que, nessa abordagem, a língua é concebida como uma rede de construções emergente do uso, de modo que a investigação linguística depende diretamente da observação de dados empíricos e, em especial, da análise da frequência de ocorrência de padrões linguísticos. Nesse contexto, a forma como os *corpora* são constituídos assume papel decisivo, uma vez que a validade das generalizações sobre uso e mudança linguística está intrinsecamente vinculada aos critérios de seleção, organização e controle dos dados.

O objetivo geral deste trabalho consiste, portanto, em discutir e propor diretrizes metodológicas para a composição de *corpora* de transcrições de vídeos do YouTube, controlados

por número de palavras, para a análise de fenômenos linguísticos no âmbito da Linguística Funcional Centrada no Uso. Para tanto, consideram-se, de modo articulado, as contribuições de Vitral (2006), no que se refere aos critérios de constituição de *corpora* e ao papel do controle quantitativo das amostras, e de Cunha Lacerda (2016), no que diz respeito à integração entre análise qualitativa e quantitativa no âmbito da abordagem construcional da mudança. Como objetivos específicos, este trabalho busca: (i) analisar as implicações da utilização de vídeos do YouTube como fonte de dados para a pesquisa em Linguística Funcional Centrada no Uso; (ii) adaptar os critérios de seleção e organização de *corpora* propostos por Vitral (2006), com ênfase no controle por número de palavras, para o contexto das transcrições de vídeos; (iii) integrar a perspectiva do método misto – qualitativo e quantitativo – à metodologia de composição e análise de *corpora*; (iv) propor um modelo de estrutura de *corpus* que permita a investigação de processos de mudança linguística e a identificação de padrões de uso em dados de fala espontânea; e (v) discutir os desafios e as potencialidades da transcrição de vídeos do YouTube para a compilação de *corpora* linguísticos.

A justificativa para a realização deste trabalho reside na constatação de uma lacuna metodológica nos estudos baseados no uso. Embora haja amplo reconhecimento do papel da frequência na modelagem da gramática e do léxico e na explicação da mudança linguística, observa-se que a discussão sobre os critérios de composição dos *corpora* – especialmente no que se refere ao controle quantitativo das amostras e à comparabilidade entre dados – ainda é pouco sistematizada. Essa lacuna torna-se ainda mais evidente diante da crescente utilização de dados provenientes de ambientes digitais, como o YouTube, cuja incorporação aos estudos linguísticos, embora promissora, exige o desenvolvimento de protocolos metodológicos específicos que assegurem rigor analítico e confiabilidade dos resultados.

A hipótese que orienta este estudo é a de que a robustez das análises baseadas em frequência – e, conseqüentemente, das generalizações sobre processos de mudança linguística – é favorecida não apenas pelos métodos analíticos empregados, mas também pelo modo como os *corpora* são constituídos e controlados. Em termos mais específicos, defende-se que, sempre que possível, a adoção de critérios sistemáticos de controle do *corpus*, especialmente no que se refere ao número de palavras, associada à integração entre análise qualitativa e quantitativa, contribui para a produção de resultados mais comparáveis e empiricamente fundamentados no âmbito da Linguística Funcional Centrada no Uso. Reconhece-se, contudo, que nem sempre o pesquisador dispõe das condições necessárias para constituir ou controlar integralmente o *corpus* utilizado.

Nesses casos, a explicitação de suas características, limitações e possíveis impactos sobre os resultados constitui requisito fundamental para a adequada interpretação das generalizações propostas.

Do ponto de vista teórico, o trabalho insere-se no âmbito das abordagens funcionalistas e dos modelos baseados no uso, assumindo, mais especificamente, os pressupostos da Linguística Funcional Centrada no Uso. Nesse domínio, dialoga particularmente com a abordagem construcional da mudança, conforme proposta por Traugott e Trousdale (2013), segundo a qual a língua se organiza em redes de construções hierarquicamente estruturadas e a frequência de uso desempenha papel relevante na emergência, convencionalização e expansão de padrões construcionais. Além disso, o estudo dialoga com a proposta de Vitral (2006), que enfatiza a importância do controle quantitativo dos *corpora*, e com a proposta de Cunha Lacerda (2016), que defende a adoção do método misto como estratégia metodológica para a análise de processos de mudança linguística.

Metodologicamente, trata-se de um estudo de natureza teórico-metodológica, que articula revisão crítica da literatura com a proposição de um modelo sistemático de constituição de *corpus*. Esse modelo fundamenta-se, de um lado, no controle quantitativo e qualitativo das amostras, contemplando não apenas o estabelecimento de *corpora* equivalentes em número de palavras, mas também, sempre que pertinente aos objetivos da investigação, o controle de variáveis como gênero textual, sequência discursiva, tipo de registro e demais características capazes de interferir na comparabilidade dos dados. De outro lado, fundamenta-se na integração entre análise qualitativa e quantitativa, de modo a possibilitar tanto a descrição contextual dos usos quanto a mensuração de sua frequência e distribuição.

No que se refere à organização do artigo, a seção 2 apresenta os fundamentos teóricos que sustentam a proposta, com destaque para a Linguística Funcional Centrada no Uso, o papel da frequência nos estudos de mudança linguística e a articulação entre análise qualitativa e quantitativa no método misto. A seção 3 discute a composição de *corpora* como um problema metodológico, evidenciando as limitações de abordagens que não consideram de forma sistemática o controle dos dados. Além disso, apresenta a proposta metodológica para a constituição de *corpora*, incluindo os princípios gerais de construção, a incorporação de dados provenientes do YouTube e a definição de um protocolo de coleta e organização dos dados. A seção 4 discute a integração entre a composição do *corpus* e o método misto, evidenciando a interdependência entre dados e análise. Por fim, a seção de Considerações Finais apresenta uma

discussão geral das implicações da proposta, indicando possíveis desdobramentos para pesquisas futuras.

2. Linguística Funcional Centrada no Uso: frequência e método misto na análise linguística

A Linguística Funcional Centrada no Uso (doravante, também, LFCU) parte do princípio de que a língua é um sistema emergente, moldado pelas experiências de uso e organizado em redes de construções que estabelecem pareamentos convencionais entre forma e função. Nesse quadro, a gramática não é concebida como um conjunto de regras abstratas e autônomas, mas como um repertório dinâmico de padrões que se constituem, se estabilizam e se transformam a partir da recorrência de usos em contextos comunicativos específicos (Martelotta, 2011; Furtado da Cunha *et al.*, 2013; Rosário; Oliveira, 2016; Bispo; Silva, 2016). Tal perspectiva implica reconhecer que a análise linguística deve estar ancorada em dados empíricos e que a observação sistemática do uso constitui condição fundamental para a descrição e explicação dos fenômenos linguísticos.

No âmbito dessa abordagem, a noção de construção assume papel central. Construções são unidades simbólicas que associam forma e significado em diferentes níveis de abstração, organizando-se em redes hierarquicamente estruturadas (Croft, 2001; Goldberg, 1995; 2006; Traugott; Trousdale, 2013). Essas redes são continuamente atualizadas por meio da experiência linguística dos falantes, de modo que padrões recorrentes tendem a se consolidar e a adquirir maior grau de convencionalização. Nesse sentido, a mudança linguística é compreendida como um processo gradual e emergente, resultante da reconfiguração dessas redes construcionais ao longo do tempo.

É nesse contexto que a frequência de uso se configura como um fator teórico fundamental. Na LFCU, a frequência não constitui apenas um dado estatístico, mas participa dos processos cognitivos e linguísticos relacionados à emergência, à convencionalização e à mudança de construções. A recorrência de determinados construtos em contextos de uso favorece processos de entrenchment cognitivo³, aumentando a acessibilidade e a estabilidade de determinados padrões no sistema linguístico. Ao mesmo tempo, a repetição desses padrões em contextos variados pode favorecer sua expansão e reconfiguração funcional, contribuindo para processos de mudança, como a construcionalização e a mudança construcional (Bybee, 2003, 2007; Traugott;

³ Neste trabalho, adota-se o termo *entrenchment* para designar o fenômeno referido na literatura de língua inglesa como *entrenchment*. Registra-se, contudo, que há variação na nomenclatura empregada em trabalhos de língua portuguesa para a tradução e a descrição desse conceito.

Trousdale, 2013; Traugott, 2022).

A literatura distingue, nesse sentido, diferentes dimensões da frequência, como a frequência *token* (relativa ao número de ocorrências de uma forma específica) e a frequência *type* (relativa à diversidade de formas que instanciam um determinado padrão). Ambas desempenham papéis distintos, mas complementares, na organização do sistema linguístico. Enquanto a frequência *token* está associada ao fortalecimento de unidades específicas e à sua automatização, a frequência *type* relaciona-se à produtividade de padrões mais abstratos, favorecendo a generalização e a formação de novos usos. Assim, a análise da frequência permite não apenas descrever a distribuição de formas linguísticas, mas também inferir processos subjacentes à sua estabilização e mudança (Bybee, 2003; 2007).

Todavia, conforme já indicado na seção anterior, a interpretação da frequência depende diretamente das condições empíricas em que os dados são observados. A frequência não possui significado analítico intrínseco; trata-se de uma medida relacional, cuja interpretação está condicionada ao universo de dados considerado. Desse modo, a utilização da frequência como evidência para processos de mudança linguística exige não apenas procedimentos quantitativos, mas também uma reflexão rigorosa sobre os critérios de constituição do *corpus*. É precisamente nesse ponto que se estabelece a articulação entre pressupostos teóricos e decisões metodológicas.

Nesse cenário, a adoção do método misto – entendido como a integração entre análise qualitativa e quantitativa – revela-se particularmente adequada aos objetivos da LFCU. Conforme argumenta Cunha Lacerda (2016), a investigação de fenômenos linguísticos, especialmente no âmbito da abordagem construcional da mudança, demanda a combinação de diferentes estratégias analíticas. Por um lado, a análise qualitativa permite identificar e descrever as propriedades formais e funcionais das construções, bem como os contextos discursivos em que ocorrem. Por outro lado, a análise quantitativa possibilita mensurar a frequência, a distribuição e a produtividade dessas construções, fornecendo evidências empíricas para a identificação de padrões de uso.

A articulação entre essas duas dimensões não deve ser entendida como uma simples justaposição de procedimentos, mas como uma integração efetiva, em que cada método informa e orienta o outro. A análise qualitativa desempenha papel fundamental na delimitação das categorias analíticas, na definição dos critérios de anotação e na interpretação dos dados, enquanto a análise quantitativa contribui para a validação empírica das generalizações propostas,

permitindo avaliar a extensão e a regularidade dos padrões identificados. Nesse sentido, o método misto constitui uma estratégia que potencializa a robustez analítica, ao combinar a sensibilidade contextual da análise qualitativa com a capacidade de generalização da análise quantitativa (Cunha Lacerda, 2016).

No âmbito da LFCU, essa integração metodológica é particularmente relevante para a investigação de processos de mudança linguística. A identificação de diferentes níveis de abstração – como construtos, microconstruções, subesquemas e esquemas (Traugott; Trousdale, 2013) – depende da análise qualitativa das relações entre forma e função, enquanto a verificação da produtividade e da expansão desses padrões exige a mensuração de sua frequência em *corpora* representativos. Assim, a articulação entre qualitativo e quantitativo não apenas enriquece a análise, mas constitui condição necessária para a compreensão dos mecanismos que governam a dinâmica do sistema linguístico.

Dessa forma, pode-se afirmar que os três eixos discutidos nesta seção – a centralidade do uso, o papel da frequência e a integração entre análise qualitativa e quantitativa – configuram o fundamento teórico-metodológico da proposta desenvolvida neste trabalho. Ao conceber a língua como um sistema emergente, cuja organização depende da recorrência de usos em contextos específicos, a LFCU atribui à frequência um papel explicativo central. Ao mesmo tempo, ao reconhecer que a análise desses padrões exige tanto descrição contextual quanto mensuração empírica, essa abordagem demanda a adoção de estratégias metodológicas integradas, como o método misto.

É a partir dessa base teórica que se justifica a ênfase, nas seções seguintes, na problematização da constituição dos *corpora* e na proposição de diretrizes metodológicas para sua organização. Em outras palavras, a discussão sobre a composição dos dados não constitui um aspecto periférico da investigação, mas uma decorrência direta dos pressupostos teóricos assumidos. Se a língua é moldada no e pelo uso e se a frequência é tomada como evidência para processos de mudança, então a forma como os dados são selecionados, organizados e controlados torna-se um elemento central para a validade das análises linguísticas.

3. A composição de *corpora* como problema metodológico e proposta de sistematização

A centralidade atribuída ao uso e à frequência na Linguística Funcional Centrada no Uso implica reconhecer que a validade das análises linguísticas depende, de maneira direta, da qualidade dos dados empíricos que as sustentam. Nesse quadro, os *corpora* deixam de ser

concebidos como simples repositórios de ocorrências e passam a constituir o fundamento sobre o qual se constroem descrições e generalizações acerca do funcionamento da língua. Apesar disso, observa-se que, em muitos estudos, a constituição do *corpus* é tratada como uma etapa instrumental, frequentemente descrita de modo sucinto, sem a devida problematização dos critérios que orientam a seleção, a organização e o controle dos dados.

Essa lacuna metodológica torna-se particularmente sensível quando se considera o papel atribuído à frequência de uso nos estudos de mudança linguística. Como amplamente reconhecido na literatura, a frequência – seja na forma de frequência *token*, seja na forma de frequência *type* – é mobilizada como evidência empírica para fenômenos como construcionalização e mudança construcional (Cunha Lacerda, 2016). No entanto, a interpretação desses valores depende necessariamente do universo de dados em que são calculados. Em outras palavras, a frequência não é um dado absoluto, mas relacional: seu significado analítico só pode ser estabelecido em função do tamanho, da composição e da natureza do *corpus* considerado.

Nesse sentido, a ausência de controle rigoroso sobre a constituição dos *corpora* pode comprometer, de maneira significativa, a validade das análises quantitativas. Um primeiro problema diz respeito à heterogeneidade no tamanho das amostras. Quando *corpora* apresentam quantidades distintas de palavras, a simples contagem de ocorrências de um determinado item torna-se insuficiente para a comparação entre conjuntos de dados. Ainda que se recorra a procedimentos de normalização estatística, a discrepância entre os tamanhos pode introduzir vieses, sobretudo quando associada a diferenças de gênero discursivo ou de contexto de uso.

Um segundo aspecto relevante refere-se à distribuição de gêneros discursivos e de tipos de registro no interior dos *corpora*. Embora essas dimensões possam apresentar correlações, elas não se confundem e devem ser consideradas separadamente na constituição e na comparação das amostras. Diferentes gêneros podem favorecer a ocorrência de determinados padrões linguísticos em função de suas características pragmáticas, discursivas e comunicativas, ao passo que diferenças de registro podem igualmente interferir nas escolhas linguísticas dos falantes, em razão de fatores como grau de formalidade, situação interacional e relação entre os participantes. Assim, *corpora* constituídos por gêneros semelhantes não necessariamente apresentam o mesmo perfil de registro, do mesmo modo que registros semelhantes podem ocorrer em gêneros distintos. Quando essas variáveis não são devidamente consideradas, corre-se o risco de atribuir à língua, de modo generalizado, padrões que podem estar associados a configurações específicas de gênero, registro ou contexto discursivo.

Além disso, a ausência de critérios explícitos de seleção e organização dos dados pode resultar em *corpora* cuja representatividade é difícil de avaliar. A noção de representatividade, embora complexa, é fundamental para que se possa sustentar que os padrões observados refletem tendências do uso linguístico e não particularidades de um conjunto restrito de dados (Traugott; Trousdale, 2013). Não há, contudo, um número universal de palavras capaz de assegurar, por si só, a representatividade de um *corpus*, uma vez que seu dimensionamento depende, entre outros fatores, das características do fenômeno investigado, de sua frequência de ocorrência, da diversidade dos contextos de uso e dos objetivos da pesquisa. Nesse sentido, o critério de saturação pode funcionar como um parâmetro complementar para a delimitação da amostra, permitindo avaliar em que medida a incorporação de novos dados continua a produzir ocorrências, padrões ou contextos analiticamente relevantes. A definição do tamanho do *corpus* deve, portanto, articular esse parâmetro a um delineamento claro de aspectos como período temporal, variedade linguística, gênero discursivo, tipo de registro e extensão das amostras, favorecendo maior comparabilidade entre estudos e maior transparência quanto às condições de produção dos resultados.

É nesse contexto que a proposta de Vitral (2006) adquire especial relevância ao enfatizar a importância do controle do número de palavras dos textos que compõem o *corpus*. Ao defender que os textos de cada conjunto analítico sejam, idealmente, de tamanho equivalente, o autor chama atenção para o fato de que a interpretação da frequência se torna metodologicamente mais consistente quando apoiada em bases empíricas comparáveis. Tal princípio contribui para reduzir a possibilidade de que diferenças observadas nos dados decorram apenas de variações no volume das amostras. Reconhece-se, contudo, que a constituição de *corpora* equivalentes nem sempre é viável, sobretudo em pesquisas diacrônicas ou em estudos que recorrem a conjuntos de dados previamente disponíveis. Nesses casos, a ausência de equivalência não inviabiliza necessariamente a análise, mas exige maior cautela na formulação de generalizações e maior transparência quanto às limitações do *corpus* e aos possíveis efeitos de sua constituição sobre os resultados.

No âmbito da Linguística Funcional Centrada no Uso, essa questão assume contornos ainda mais relevantes quando se considera a articulação entre análise qualitativa e quantitativa, tal como proposta pelo método misto, conforme assumido por Cunha Lacerda (2016). Como destaca a autora, se, por um lado, a análise qualitativa permite a identificação e a descrição das construções em seus contextos de uso, por outro, a análise quantitativa depende de uma base de

dados estruturada de forma consistente para que seja possível mensurar frequência, produtividade e distribuição. Desse modo, a eficácia do método misto está intrinsecamente vinculada à forma como o *corpus* é constituído, o que reforça a necessidade de tratar sua composição como uma etapa central da metodologia.

O problema da constituição de *corpora* torna-se ainda mais evidente no contexto contemporâneo, marcado pela ampla disponibilidade de dados linguísticos em ambientes digitais. Plataformas como o YouTube oferecem acesso a uma quantidade expressiva de produções discursivas, muitas das quais caracterizadas por alto grau de espontaneidade e diversidade sociolinguística. Embora esses dados representem uma oportunidade valiosa para os estudos baseados no uso, sua incorporação sem critérios rigorosos de seleção e controle pode acentuar os problemas já mencionados, ampliando a heterogeneidade e dificultando a comparabilidade das análises.

Diante desse cenário, torna-se necessário deslocar o estatuto do *corpus* no interior da pesquisa linguística: de um recurso operacional para um componente metodológico estruturante. Isso implica reconhecer que decisões relativas à seleção, à organização e ao controle dos dados não são neutras, mas influenciam diretamente os resultados obtidos e as conclusões formuladas. Em termos mais precisos, pode-se afirmar que, na medida em que a frequência é tomada como evidência empírica para a descrição e explicação da língua, a composição do *corpus* passa a constituir uma variável metodológica determinante.

É nesse sentido que se propõe, neste trabalho, um modelo metodológico para a constituição de *corpora* orais baseados em transcrições de vídeos do YouTube. Tal proposta busca sistematizar critérios que assegurem comparabilidade, representatividade e adequação dos dados aos objetivos analíticos, articulando três eixos fundamentais: (i) princípios gerais de construção do *corpus*, (ii) incorporação de dados provenientes de ambientes digitais e (iii) definição de um protocolo sistemático de coleta e organização dos dados.

No que se refere aos princípios gerais de construção do *corpus*, destaca-se, em primeiro lugar, a importância do controle quantitativo das amostras, especialmente quando a investigação envolve a comparação entre diferentes conjuntos de dados. Dependendo dos objetivos da pesquisa, pode ser necessário organizar o material em diferentes *subcorpora* ou mesmo constituir *corpora* distintos, delimitados, por exemplo, segundo gênero discursivo, período histórico, variedade linguística, perfil sociolinguístico ou contexto comunicativo. Assim, em uma investigação que compare diferentes gêneros, pode-se constituir, por exemplo, um *subcorpus* de vlogs e outro

de debates, buscando-se, sempre que possível, controlar a extensão de cada conjunto e outros parâmetros relevantes para a análise. A definição prévia do volume de dados – como o estabelecimento de um número equivalente de palavras para cada conjunto – favorece a comparabilidade e oferece uma base mais consistente para a interpretação dos resultados quantitativos. Esse controle, contudo, deve ser articulado a outros critérios de constituição das amostras, uma vez que a equivalência em número de palavras, isoladamente, não garante sua comparabilidade. A representatividade dos dados deve, portanto, ser avaliada a partir de parâmetros definidos em consonância com os objetivos da pesquisa, considerando-se, entre outros aspectos, gênero discursivo, tipo de registro, variedade linguística, perfil sociolinguístico e contexto comunicativo.

A incorporação de dados provenientes do YouTube configura-se como um elemento inovador dessa proposta metodológica. Essa plataforma oferece acesso a uma ampla variedade de dados de fala espontânea ou semi-espontânea, distribuídos em diferentes gêneros discursivos, como entrevistas, *vlogs*, *podcasts* e debates. Tal diversidade permite a observação de fenômenos linguísticos em contextos próximos à interação natural, ampliando as possibilidades de análise no âmbito da Linguística Funcional Centrada no Uso. No entanto, a utilização desses dados exige a definição de critérios rigorosos de seleção, considerando aspectos como grau de espontaneidade, qualidade do áudio, ausência de sobreposição excessiva de fala e adequação ao objeto de estudo. A heterogeneidade dos dados digitais demanda, portanto, procedimentos sistemáticos que garantam sua organização e comparabilidade.

Nesse contexto, propõe-se um protocolo de coleta e constituição de *corpus* que articula seleção criteriosa dos dados, controle quantitativo das amostras e preparação sistemática para análise. O protocolo inicia-se com a delimitação do fenômeno linguístico a ser investigado, assegurando a coerência entre os dados coletados e os objetivos da pesquisa. Em seguida, estabelece-se a definição dos critérios de seleção dos vídeos, contemplando gênero discursivo, grau de espontaneidade, perfil sociolinguístico e qualidade técnica. A estratégia de amostragem baseia-se no controle do número de palavras, de modo que cada *subcorpus* seja constituído por um volume equivalente de dados, garantindo comparabilidade.

A etapa de coleta envolve a seleção dos vídeos e o registro sistemático de informações relevantes, como título, canal, *link*, data de publicação e características do conteúdo. Posteriormente, realiza-se a transcrição dos dados, adotando-se um modelo ortográfico adaptado que preserve traços relevantes da fala, como hesitações e repetições. Em seguida, procede-se à

limpeza e segmentação dos dados, com a exclusão de trechos inaudíveis ou irrelevantes e a organização em unidades analisáveis. O controle quantitativo é, então, realizado por meio da contagem do número de palavras e do ajuste das amostras até atingir o volume previamente definido.

Por fim, realiza-se a codificação das ocorrências do fenômeno linguístico investigado, acompanhada da anotação de informações contextuais relevantes, bem como a validação do *corpus*, mediante a verificação da consistência dos critérios adotados e da identificação de possíveis vieses. Esse conjunto de procedimentos visa a favorecer uma interpretação mais confiável da frequência observada nos dados e de sua relação com os padrões de uso linguístico. A título de ilustração, considere-se uma investigação hipotética que tenha como objetivo comparar a frequência e a distribuição de determinado fenômeno linguístico em diferentes gêneros discursivos. Nesse caso, o pesquisador poderia organizar o *corpus* em diferentes *subcorpora*, constituindo, por exemplo, um conjunto de dados provenientes de vlogs, outro de debates e outro de entrevistas. Para favorecer a comparabilidade, poderia estabelecer, sempre que as características dos dados assim o permitissem, uma extensão equivalente para cada *subcorpus* — por exemplo, 50 mil palavras —, além de controlar outros parâmetros relevantes, como período de produção, variedade linguística, tipo de registro e perfil sociolinguístico dos participantes. As ocorrências do fenômeno investigado seriam, então, identificadas e codificadas em cada *subcorpus* segundo fatores linguísticos e contextuais previamente definidos. Esse procedimento permitiria comparar não apenas a frequência relativa do fenômeno nos diferentes conjuntos, mas também os contextos em que ele ocorre e os fatores que favorecem ou restringem seu uso. Nesse exemplo, portanto, a equivalência no número de palavras constitui apenas um dos parâmetros de controle, devendo ser articulada a outros critérios definidos em função da pergunta de pesquisa e das características do fenômeno analisado.

4. Integração entre composição de *corpus* e método misto

A discussão desenvolvida nas seções anteriores permite afirmar que a constituição do *corpus* não pode ser concebida como uma etapa prévia e independente da análise linguística, mas como um componente metodológico intrinsecamente articulado às estratégias analíticas adotadas. Nesse sentido, a integração entre a composição do *corpus* e o método misto — entendido como a articulação entre abordagens qualitativas e quantitativas — configura-se como

um eixo central para a investigação de fenômenos linguísticos no âmbito da Linguística Funcional Centrada no Uso, tal como assume Cunha Lacerda (2016).

No quadro teórico assumido neste trabalho, a língua é concebida como uma rede de construções emergente do uso, cuja descrição e explicação dependem da observação de dados empíricos e da análise de padrões recorrentes. Nesse contexto, como argumenta Cunha Lacerda (2016), a adoção do método misto constitui uma estratégia importante, uma vez que permite articular a análise qualitativa dos contextos de uso à mensuração quantitativa da frequência de ocorrência das construções. A análise qualitativa possibilita a identificação das propriedades formais e funcionais das construções, bem como a descrição dos ambientes discursivos em que emergem, enquanto a análise quantitativa fornece evidências empíricas sobre sua produtividade, distribuição e possível entrenchamento no sistema linguístico.

No entanto, a eficácia dessa articulação metodológica está diretamente condicionada à forma como o *corpus* é constituído. Como destaca Vitral (2006), a interpretação dos dados de frequência depende da relação entre o número de ocorrências de um determinado item e o universo total de palavras analisadas. Dessa forma, sem um controle rigoroso do *corpus* – especialmente no que se refere ao número de palavras e à comparabilidade entre os conjuntos de dados –, a análise quantitativa perde sua força explicativa, uma vez que as frequências observadas podem refletir apenas diferenças no volume das amostras, e não padrões reais de uso.

A integração entre composição de *corpus* e método misto pode ser compreendida, portanto, como uma relação de interdependência. Por um lado, a análise qualitativa orienta a constituição do *corpus*, na medida em que define os critérios de seleção dos dados, os contextos relevantes e as categorias analíticas a serem consideradas. Por outro lado, a análise quantitativa exige um *corpus* estruturado de forma sistemática, que permita a mensuração confiável da frequência e a comparação entre diferentes contextos de uso. Nesse sentido, a constituição do *corpus* não apenas precede a análise, mas participa ativamente de sua construção.

No âmbito da abordagem construcional da mudança, essa interdependência assume um papel ainda mais relevante. Como evidencia Cunha Lacerda (2016), a identificação de diferentes níveis de abstração – como construto, microconstrução, subesquema e esquema (Traugott; Trousdale, 2013) – depende da articulação entre descrição qualitativa e levantamento quantitativo das ocorrências. A análise qualitativa permite caracterizar o pareamento entre forma e função e identificar padrões recorrentes, enquanto a análise quantitativa possibilita mensurar a extensão desses padrões e verificar sua produtividade no sistema linguístico. No entanto, essa

mensuração só é possível quando o *corpus* apresenta condições adequadas de comparabilidade, o que reforça a importância do controle quantitativo defendido por Vitral (2006).

Isso significa que comparar *corpora* díspares – isto é, com extensões significativamente diferentes – não compromete apenas a comparação numérica direta, mas afeta, de modo mais profundo, a própria interpretação linguística dos dados. Em termos metodológicos, a frequência de um item não é um valor absoluto, mas uma relação entre o número de ocorrências e o total de palavras do *corpus*. Quando essa base de comparação não é controlada, a frequência deixa de refletir padrões de uso e passa a refletir, em grande medida, diferenças estruturais entre os *corpora*. Nesse sentido, como argumenta Vitral (2006), a análise da frequência só pode ser interpretada de forma consistente quando relacionada a uma base empírica comparável, o que pressupõe o controle do número de palavras dos textos que compõem o *corpus*.

Em primeiro lugar, *corpora* maiores tendem a apresentar maior número de ocorrências de praticamente qualquer item, o que pode criar a falsa impressão de maior produtividade ou relevância funcional de determinada construção. Ainda que se recorra a medidas de frequência relativa – por exemplo, ocorrências por mil palavras –, a discrepância de tamanho pode continuar produzindo distorções, especialmente quando associada à distribuição desigual de contextos discursivos. Isso ocorre porque *corpora* mais extensos tendem a abarcar maior diversidade de situações comunicativas, aumentando a probabilidade de ocorrência de usos menos frequentes ou mais especializados, o que não necessariamente reflete diferenças estruturais na língua, mas sim diferenças na cobertura amostral.

Em segundo lugar, a comparação entre *corpora* de tamanhos distintos dificulta a identificação de padrões de distribuição. Uma construção que aparece com baixa frequência em um *corpus* menor pode ser interpretada como marginal ou pouco produtiva, quando, na verdade, sua baixa ocorrência decorre da limitação da amostra. Por outro lado, em um *corpus* maior, essa mesma construção pode apresentar um número absoluto mais elevado de ocorrências, sugerindo maior relevância, ainda que sua frequência relativa permaneça estável. Sem controle do tamanho dos *corpora*, torna-se difícil distinguir entre efeitos de amostragem e propriedades do sistema linguístico.

Além disso, a ausência de equivalência quantitativa compromete a análise comparativa entre diferentes contextos – por exemplo, entre gêneros discursivos, variedades linguísticas ou períodos históricos. Como destaca Vitral (2006), a comparação entre *corpora* só é metodologicamente válida quando se estabelece uma base comum de referência, o que implica,

idealmente, a equivalência no número de palavras. Sem essa equivalência, diferenças observadas podem ser atribuídas indevidamente a fatores linguísticos, quando, na realidade, decorrem de diferenças na dimensão dos dados.

No âmbito da Linguística Funcional Centrada no Uso, esse problema torna-se ainda mais crítico, uma vez que a frequência é frequentemente mobilizada como evidência para processos como produtividade e mudança construcional. Como argumentam Traugott e Trousdale (2013), a organização da língua em redes construcionais hierarquicamente estruturadas pressupõe a identificação de padrões recorrentes de uso, cuja frequência desempenha papel central na emergência e na consolidação das construções. De modo convergente, Cunha Lacerda (2016) destaca que a articulação entre análise qualitativa e quantitativa – no âmbito do método misto – é fundamental para que tais padrões sejam descritos e empiricamente validados. No entanto, essa validação depende de dados quantitativamente comparáveis; caso contrário, as inferências sobre frequência e produtividade tornam-se metodologicamente frágeis.

Por exemplo, pode-se interpretar como aumento de produtividade o que, na verdade, é apenas resultado de um *corpus* mais extenso ou mais diversificado. Nesse caso, o que aparece como evidência de mudança linguística pode ser apenas um efeito de amostragem, comprometendo a análise dos processos de construcionalização e mudança construcional descritos por Traugott e Trousdale (2013).

Por fim, é importante destacar que o problema não se limita ao tamanho absoluto dos *corpora*, mas envolve também a sua estrutura interna. *Corpora* maiores e não controlados tendem a incluir proporções distintas de gêneros, registros e contextos, o que reforça o viés analítico. Nesse sentido, o controle do número de palavras – conforme proposto por Vitral (2006) – deve ser entendido como parte de um conjunto mais amplo de procedimentos metodológicos voltados à garantia da comparabilidade dos dados.

Assim, a comparação entre *corpora* díspares não apenas dificulta a análise quantitativa, mas compromete a própria validade das generalizações linguísticas. É por essa razão que o controle do tamanho das amostras – aliado à integração entre análise qualitativa e quantitativa, conforme defendido por Cunha Lacerda (2016), e à compreensão do papel da frequência nos processos de mudança, conforme Traugott e Trousdale (2013) – deve ser considerado um princípio fundamental na constituição de *corpora* para estudos baseados no uso.

5. Considerações finais

Este artigo teve como propósito discutir e propor diretrizes metodológicas para a constituição de *corpora* orais, com foco na utilização de transcrições de vídeos do YouTube no âmbito da Linguística Funcional Centrada no Uso. Partiu-se do reconhecimento de que a validade das análises baseadas em frequência – elemento central nessa abordagem – depende diretamente dos critérios que orientam a seleção, a organização e o controle dos dados empíricos.

No que se refere ao cumprimento dos objetivos inicialmente propostos, considera-se que foram plenamente alcançados. O objetivo geral – propor um modelo metodológico para a composição de *corpora* controlados por número de palavras – foi desenvolvido a partir da articulação entre a problematização da constituição dos dados e a formulação de diretrizes sistemáticas para sua organização. Da mesma forma, os objetivos específicos foram contemplados ao longo do trabalho: analisaram-se as implicações do uso de vídeos do YouTube como fonte de dados, adaptaram-se os critérios de constituição de *corpus* propostos por Vitral (2006) ao contexto de dados digitais, integrou-se a perspectiva do método misto, conforme proposta por Cunha Lacerda (2016), e propôs-se um protocolo metodológico que permite a investigação de padrões de uso em dados de fala espontânea. Ademais, foram discutidos os principais desafios e potencialidades envolvidos na transcrição e organização de dados provenientes do YouTube.

Quanto à hipótese que orientou este estudo – a de que a validade das análises baseadas em frequência depende, fundamentalmente, da forma como os *corpora* são constituídos –, considera-se que foi confirmada no plano teórico-metodológico. A análise desenvolvida evidenciou que a ausência de controle sistemático dos dados compromete a interpretação das frequências e pode gerar inferências inadequadas sobre padrões linguísticos. Em contrapartida, a adoção de critérios rigorosos de constituição de *corpus*, especialmente no que se refere ao controle do número de palavras, conforme defendido por Vitral (2006), associada à integração entre análise qualitativa e quantitativa, conforme proposto por Cunha Lacerda (2016), configura-se como condição essencial para a produção de resultados mais robustos e empiricamente fundamentados. Nesse sentido, a discussão também se articula à perspectiva de Traugott e Trousdale (2013), ao evidenciar que a identificação de padrões construcionais depende da observação sistemática da frequência em dados comparáveis.

As contribuições deste trabalho para a área situam-se, sobretudo, no plano metodológico. Em primeiro lugar, o artigo contribui ao reposicionar a composição do *corpus* como um elemento

estruturante da investigação linguística, e não como uma etapa meramente operacional. Em segundo lugar, oferece uma proposta sistematizada de constituição de *corpora* orais baseados em dados digitais, incorporando o controle quantitativo das amostras como princípio central. Em terceiro lugar, promove a integração entre diferentes tradições metodológicas – em particular, a proposta de controle de *corpus* de Vitral (2006) e o método misto na abordagem construcional (Cunha Lacerda, 2016) –, contribuindo para o fortalecimento de uma abordagem empiricamente orientada na Linguística Funcional Centrada no Uso. Por fim, ao explorar o potencial do YouTube como fonte de dados linguísticos, o trabalho amplia o escopo empírico da área e aponta caminhos para a incorporação sistemática de dados digitais na pesquisa linguística.

No que diz respeito a trabalhos futuros, diversas possibilidades se abrem a partir desta proposta. Em primeiro lugar, destaca-se a necessidade de aplicação empírica do modelo metodológico aqui proposto, com a constituição de *corpora* específicos e a análise de fenômenos linguísticos concretos, o que permitirá avaliar, em termos práticos, a eficácia das diretrizes estabelecidas. Em segundo lugar, investigações futuras podem aprofundar a discussão sobre critérios de representatividade e balanceamento de *corpora* em ambientes digitais, considerando variáveis sociolinguísticas e discursivas mais refinadas. Além disso, há espaço para o desenvolvimento de ferramentas automatizadas de transcrição e anotação que possam otimizar a constituição de *corpora* orais a partir de vídeos. Por fim, estudos comparativos entre *corpora* tradicionais e *corpora* digitais poderão contribuir para uma compreensão mais ampla das implicações metodológicas do uso de diferentes tipos de dados na Linguística Funcional Centrada no Uso.

Em síntese, ao enfatizar a interdependência entre dados e análise, este trabalho reafirma que a robustez das investigações linguísticas depende, em larga medida, do rigor metodológico na constituição dos *corpora*. Nesse sentido, espera-se que as reflexões aqui apresentadas possam contribuir para o aprimoramento das práticas de pesquisa na área e para o desenvolvimento de estudos cada vez mais consistentes e empiricamente fundamentados.

Referências

- BISPO, E. B.; SILVA, J. R. Variação linguística, mudança linguística e construcionalização. In: *XXI Seminário do Grupo de Estudos Discurso & Gramática e VIII Seminário Internacional do Grupo de Estudos Discurso & Gramática*. Rio de Janeiro: UFRJ, 2016.
- BYBEE, J. Mechanisms of change in grammaticization: the role of frequency. In: JOSEPH, B. D.; JANDA, R. D. (org.). *The handbook of historical linguistics*. Oxford: Blackwell, 2003. p. 602–623.

BYBEE, J. *Frequency of use and the organization of language*. Oxford: Oxford University Press, 2007.

CROFT, W. *Radical construction grammar: syntactic theory in typological perspective*. Oxford: Oxford University Press, 2001.

CUNHA LACERDA, P. F. A. da. O papel do método misto na análise de processos de mudança em uma abordagem construcional: reflexões e propostas. *Revista Linguística / Revista do Programa de Pós-Graduação em Linguística da Universidade Federal do Rio de Janeiro*. Volume Especial, dez de 2016, p. 83-101.

FURTADO DA CUNHA, M. A.; BISPO, E. B.; SILVA, J. R. Linguística funcional centrada no uso: conceitos básicos e categorias analíticas. In: CEZÁRIO, M. M.; FURTADO DA CUNHA, M. A. (Orgs.). *Linguística centrada no uso: uma homenagem a Mário Martelotta*. Rio de Janeiro: Mauad, 2013, p. 13-44.

GOLDBERG, A. E. *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press, 1995.

GOLDBERG, A. E. *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press, 2006.

MCENERY, T.; HARDIE, A. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press, 2012.

ROSÁRIO, I. da C. do; OLIVEIRA, M. R. de. Funcionalismo e abordagem construcional da gramática. *Alfa*, São Paulo, 60 (2), p. 233-259, 2016.

TRAUGOTT, E. C.; TROUSDALE, G. *Constructionalization and constructional changes*. Oxford: Oxford University Press, 2013.

VITRAL, L. O papel da frequência na identificação de processos de gramaticalização. *Scripta*, vol. 9, n. 18. Belo Horizonte, 2006, p. 149-177.
