Revista Brasileira de Pesquisa em Educação em Ciências
doi: 10.28976/1984-2686rbpec2020u799822

Concepções de *Corpus* de Análise na Pesquisa em Educação em Ciências Naturais: Uma Investigação em Dissertações e Teses de um Programa de Pós-Graduação

Conceptions of *Corpus* for Analysis in Science Education Research: An Investigation in Dissertations and Theses from a Graduate Program

Concepciones sobre el *Corpus* de Análisis en la Investigación en Educación Científica: Una Búsqueda Científica en Disertaciones y Tesis de un Programa de Posgrado

Julio Murilo Trevas dos Santos Brasil
Neide Maria Michelan Kiouranis Brasil

Investigaram-se neste trabalho compreensões e concepções sobre *corpus* de análise em dissertações e teses de um programa de pós-graduação em Educação em Ciências e Matemática. Buscou-se então: avaliar a ocorrência dessas compreensões e categorizá-las; identificar polissemia em relação ao termo *corpus*; avaliar as compreensões em função de referenciais teóricos de análise textual. As dissertações e as teses foram integralmente organizadas em *corpora*. Processaram-se os *corpora* por meio de análise textual auxiliada por computador (CATA), em um processo que envolveu métodos quantitativos e qualitativos. Utilizou-se o servidor Voyant Tools, um ambiente Web de análise de *corpora* extensos. A análise dos *corpora* resultou em quatro categorias a posteriori: caracterização do *corpus*; definição do *corpus*; formação do *corpus*; e operação sobre o *corpus*. Embora os resultados indiquem que as compreensões estavam vinculadas às metodologias de análise textual, perceberam-se inconsistências, afastamentos de referenciais teóricos e a necessidade de critérios para a constituição de *corpora*.

Palavras-chave: Análise de *Corpora*; Análise de Conteúdo; Análise Textual.

In this work, we investigate the understandings and conceptions of *corpus* for analysis in dissertations and theses of a graduate program in Science and Mathematics Education. Our goal is to evaluate the occurrence of these understandings and categorize them; to identify polysemy in relation to the term "*corpus*"; to evaluate understandings according to benchmarks of theoretical textual analysis. The dissertations and theses were organized

entirely in *corpora*, which were processed through computer-aided textual analysis (CATA) in a process that involved quantitative and qualitative methods. For this, we used the Voyant Tools server, a web-based environment for extensive *corpora* analysis. The results pointed to four a posteriori categories: *corpus* characterization, definition formation and operation. Although the results indicate that the understandings were linked to methodologies of textual analysis, we found inconsistencies, deviations from theoretical benchmarks and the need for criteria for *corpora* development.

Keywords: *Corpora* Analysis; Content Analysis; Textual Analysis.

En este trabajo se investigó las comprensiones y concepciones sobre el *corpus* de análisis en disertaciones y tesis de un programa de posgrado en Educación en Ciencias y Matemáticas. El objetivo era evaluar la ocurrencia de estos entendimientos y clasificarlos; identificar polisemia en relación con el término *corpus*; evaluar las comprensiones en función de las referencias teóricas del análisis textual. Las disertaciones y tesis fueron totalmente organizadas en *corpora*. Los *corpora* se procesaron mediante análisis textual asistido por computadora (CATA), en un proceso que involucraba métodos cuantitativos y cualitativos. Se utilizó el servidor Voyant Tools, un entorno web de análisis de *corpora* extenso. El análisis de los *corpora* resultó en cuatro categorías a posteriori: caracterización del *corpus*; definición del *corpus*; formación del *corpus*; y operación en el *corpus*. Aunque los resultados indican que los entendimientos estaban vinculados a metodologías de análisis textual, se percibieron inconsistencias, desviaciones de las referencias teóricas y la necesidad de criterios para la constitución de *corpus*.

Palabras clave: Análisis de *Corpora*; Análisis de Contenido; Análisis Textual.

Introdução

Metodologias e técnicas de análise textual são muito utilizadas nas pesquisas na área de Educação em Ciências Naturais (Oliveira et al. 2003; Pinhão & Martins, 2009; Santos et al., 2018). De um modo geral, três análises se destacam: a Análise de Conteúdo (AC); a Análise de Discurso (AD); e a Análise Textual Discursiva (ATD) (Chrysostomo & Messeder, 2017; Santos et al., 2017). Essas análises se diferenciam em seus objetivos, mas mantêm alguns elementos comuns. Conforme discussão de Rocha e Deusdará (2005), existem aproximações e afastamentos entre essas análises. O foco deste trabalho está em um desses elementos comuns que pode estabelecer aproximação: o *corpus* de análise.

A preocupação com *corpus* é consequência da preocupação com a qualidade da análise e da pesquisa em que ela está inserida. A qualidade mencionada não se refere ao dualismo (Gamboa, 2003) pesquisa qualitativa e pesquisa quantitativa, mas às características que tornam a pesquisa reconhecida e aceita pelo meio acadêmico. Neste trabalho, se adota uma perspectiva de pesquisa qualitativa de métodos mistos, ou seja, que pode in*corpora*r tanto métodos qualitativos, quanto métodos quantitativos.

Contudo, admite-se que as discussões do dualismo permeiam as concepções e os critérios de qualidade.

Se existe uma relação entre *corpus* e a qualidade, quer seja da análise ou da pesquisa, surgem as dúvidas se um *corpus* deve ser compreendido como um critério para uma análise e se deve ser elaborado segundo critérios. Bauer e Gaskell (2015, p. 478) por exemplo, criticaram a representatividade, a fidedignidade, a validade e a amostragem como critérios de qualidade da pesquisa qualitativa. Esses autores propuseram "critérios com equivalência funcional à tradição quantitativa" (Bauer & Gaskell, 2015, p. 480), dentre os quais consta a "construção do *corpus*" (Bauer & Gaskell, 2015, p. 481) como equivalente à validade interna. Orlandi (2015, p. 60), como outro exemplo, não caracterizou um *corpus* como critério de qualidade de uma pesquisa. A autora afirmou que uma das primeiras etapas de uma AD é a constituição e a delimitação, segundo critérios teóricos, de um *corpus*. Complementando, Bauer e Aarts (2015, p. 61) e Bardin (2011, p. 126) apresentaram passos e regras para a constituição de um *corpus*.

A dúvida sobre *corpus* e as diferentes abordagens nos exemplos citados sugeriram uma pergunta: como são as compreensões e/ou concepções de pesquisadores sobre *corpus* para uma análise textual? Algumas suposições para a resposta foram: compreensão de *corpus* vinculada à metodologia de análise adotada; uma concepção de *corpus* comum às metodologias de análise; concepções de *corpus* desvinculadas de teorias e metodologias de análise, baseadas em senso comum ou representações de comunidades de pesquisa; falta de compreensão ou concepção sobre o *corpus*.

Efetivamente, a pergunta supracitada surgiu no desenvolvimento de uma pesquisa que concilia métodos mistos de análise textual. Tornou-se uma necessidade na mencionada pesquisa, dirimir a dúvida sobre o *corpus* para permitir o estabelecimento de alguns requisitos de confiabilidade e validade da abordagem analítica adotada. Sabe-se que as análises textuais são realizadas por muitos pesquisadores de várias áreas de conhecimento. É inviável em um curto período de tempo realizar uma investigação envolvendo um número representativo (abrangência regional ou nacional) desses pesquisadores. No intuito de tentar responder a pergunta e considerando: a) que este trabalho se configurou como etapa de uma pesquisa; b) a investigação envolvendo número expressivo de pesquisadores não é objetivo da pesquisa; c) o interesse nas pesquisas em Educação em Ciências Naturais; o campo de investigação deste trabalho foi delimitado a um programa de pós-graduação: Programa de Pós-Graduação em Educação para a Ciência e a Matemática (PCM) da Universidade Estadual de Maringá (UEM), no campus de Maringá. Este é o programa ao qual os autores deste trabalho estão vinculados. Trata-se de um dos programas mais antigos na área de Ensino de Ciências e Matemática no Paraná, tendo inciado atividades em 2004 (a maioria dos programas iniciou atividades após 2012), e um dos quatro a ofertar doutorado acadêmico. O PCM oferta capacitação stricto sensu nos níveis de mestrado e doutorado (acadêmicos) e está estruturado em três linhas de pesquisa: "Ensino e Aprendizagem na Educação Científica"; "Formação de Professores de Ciências e Matemática"; "História, Epistemologia e Cultura da Ciência".

O programa possui um corpo docente que se renovou ao longo dos anos, mas que se identifica pertencente a quatro principais áreas: Ciências Biológicas, Física, Matemática e Química. Alguns docentes orientam apenas em nível de mestrado, enquanto outros orientam nos dois níveis. Tendo em vista as características do programa, considerou-se que o campo delimitado apresenta certa representatividade. Buscou-se então, nas dissertações e teses defendidas nesse programa, os indicadores para a proposição de uma resposta (não generalizável) à pergunta.

Um programa de pós-graduação é um campo de investigação importante. Primeiro porque a pós-graduação é um locus natural de pesquisa. Segundo porque estabelece relações entre pesquisadores formadores e pesquisadores em formação. Terceiro porque há um efeito multiplicador, ou seja, egressos de um programa formam novos núcleos de pesquisas. E nesses núcleos, as pesquisas são conduzidas conforme os pressupostos teóricos e metodológicos do programa de pós-graduação. Uma quarta justificativa da importância é a preocupação, em um programa de pós-graduação, com paradigmas de pesquisa atualizados. Outra justificativa é o compartilhamento de experiências entre os pesquisadores de diferentes programas. Resumidamente, uma pós-graduação apresenta ramificação e capilaridade na pesquisa. Por isso os indicadores obtidos desse campo de investigação permitem uma previsão sobre campos mais amplos. Destaca-se também que esses indicadores permitirão discutir: critérios para a constituição de *corpus*; e o *corpus* como critério para uma análise e para a qualidade de uma pesquisa.

Portanto o objetivo deste trabalho foi identificar como se apresentam as compreensões e concepções de *corpus* de análise nas dissertações e teses do PCM. Para alcançar esse objetivo geral, foram elencados alguns objetivos específicos: avaliar a ocorrência do termo *corpus*; categorizar as ocorrências identificadas; avaliar uma possível polissemia em relação ao termo *corpus*; identificar convergências e divergências entre as concepções; avaliar compreensões e concepções em função do referencial teórico adotado.

Referencial Teórico

Para que fossem identificadas e discutidas as compreensões e concepções em dissertações e teses, foi necessário resgatar da literatura pertinente os conceitos de *corpus* das principais propostas de análise textual. *Corpus* é termo latino, cujo plural é *corpora*, que originalmente significa corpo (Glosbe, 2020; Sardinha, 2000). O dicionário Priberam (Informática, 2013), por exemplo, define *corpus* como "coletânea acerca de um mesmo assunto" e como "conjunto de documentos que servem de base para a descrição ou o estudo de um fenômeno". O conceito de *corpus* como coletânea não é recente. Há registros da produção de *corpora* de citações da Bíblia na Antiguidade e Idade Média (Sardinha, 2000). Porém, a noção de *corpus* para análise é proveniente da área de Linguística no século XX (Sardinha, 2000). Segundo Aluísio e Almeida (2006), existem duas grandes perspectivas para *corpus*: uma da Linguística e outra da Linguística de *Corpus* (LC). Na perspectiva da Linguística, o *corpus* é um conjunto finito e variado de

enunciados em uma língua, o qual é tomado como objeto de análise (Sardinha, 2000). Na perspectiva da LC, Sardinha (2000) cita como definição mais completa a elaborada por Sánchez (1995): "Um conjunto de dados lingüísticos [...], sistematizados segundo determinados critérios, [...] dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise". A perspectiva da LC se destaca da Linguística por exigir um material textual processável por computador. Isso é compreensível porque a LC se dedica à análise de *corpora* extensos, com milhões de palavras (Aluísio & Almeida, 2006; Sardinha, 2000).

Em um campo com proximidade à Linguística, Eni Orlandi (2015, p. 60) discutiu *corpus* em uma AD (neste trabalho, análise de discurso de linha francesa). Constituir o *corpus* com materiais de análise é uma das primeiras etapas da AD, cujas práticas discursivas podem envolver letra, imagem, som, entre outras. A autora mencionou dois tipos de *corpus*: o experimental; e o de arquivo. Quando os materiais de análise são produzidos especificamente para a pesquisa, o *corpus* é dito experimental (ou empírico). Aiub (2012, p. 72) comentou que "[…] os *corpora* experimentais são baseados no que podemos chamar, a grosso modo, de coleta de dados [...]". Quando são materiais preexistentes, como livros ou documentos, o *corpus* é dito de arquivo (Caregnato & Mutti, 2006). Uma outra distinção efetuada por Orlandi refere-se a texto e discurso. O texto "[…] é a unidade que o analista tem diante de si e do qual ele parte", enquanto o discurso "[…] se explicita em suas regularidades pela sua referência a uma ou outra formação discursiva" (Orlandi, 2015, p. 61). Então a autora propôs o *corpus* como construção de "montagens discursivas que obedecem critérios que decorrem de princípios teóricos da análise de discurso" (Orlandi, 2015, p. 61).

No campo da Educação em Ciências Naturais, Moraes e Galiazzi (2011, p. 16) indicaram que *corpus*, na ATD, foi uma denominação retirada da obra de Bardin (2011). Eles afirmaram que *corpus* é um conjunto de documentos, "essencialmente produções textuais" (Moraes, 2003), através do qual toda análise textual se concretiza. Os textos, por sua vez, são produções linguísticas que expressam discursos sobre um fenômeno em um período determinado. Essas produções possuem sentido mais amplo abrangendo "imagens e outras expressões linguísticas" (Moraes & Galiazzi, 2011, p. 16), podendo ser produzidas para a análise ou selecionadas de documentos previamente existentes. Noções essas que coincidem com o que foi defendido por Orlandi na AD. Moraes assumiu não trabalhar com todo o *corpus*, mas defendeu a necessidade de definir uma amostra obtida "de um conjunto maior de textos" (Moraes, 2003). Enquanto na Linguística o *corpus* é o material analisado, para Moraes o material analisado é um excerto do *corpus*.

Corpus também foi discutido por Bardin (2011) na sua proposição de AC (análise de conteúdo de linha francesa). Para a autora, um *corpus* "[…] é o conjunto dos documentos tidos em conta para serem submetidos aos procedimentos analíticos" (Bardin, 2011, p. 126). Desse modo, um único documento não constitui um *corpus* (Bardin, 2011, p. 128). Bardin mencionou que a constituição do *corpus* "[…] implica, muitas vezes, escolhas, seleções e regras" (Bardin, 2011, p. 126). Resumidamente,

algumas regras (Bardin, 2011, p. 126) são: a) os documentos selecionados devem ser adequados à análise do fenômeno investigado; b) os documentos devem "[…] obedecer critérios precisos de escolha" (Bardin, 2011, p. 128); c) todos os documentos que atendem o campo e os critérios do *corpus* devem ser incluídos; d) os documentos devem ser representativos do universo de materiais analisáveis.

Na abordagem estadunidense de AC, que é a perspectiva fundadora na área (Carlomagno & Rocha, 2016), Neuendorf (2017, p. 80) estabeleceu uma relação entre arquivo e *corpus*. Em sua abordagem declaradamente quantitativa, as análises são realizadas sobre arquivos. Arquivo é "[…] uma coleção de mensagens, geralmente bem indexadas" (Neuendorf, 2017, p. 80), armazenadas em formato eletrônico. Nesse caso, uma concordância com a LC pela necessidade do formato digital. O formato eletrônico permite um acesso mais fácil e operações complexas que não eram possíveis antes do uso de computadores. Em áreas das Ciências Humanas e Sociais, os arquivos são frequentemente denominados *corpora*. Nesse contexto, um *corpus* é "[…] tipicamente um conjunto de materiais escritos representando uma época e um lugar em particular" (Neuendorf, 2017, p. 81). Neuendorf mencionou que foram estabelecidos padrões para o armazenamento e transferência de textos, que significam critérios para os *corpora*.

Bauer e Aarts (2015) também discutiram sobre o *corpus*. Eles apresentaram algumas noções de *corpus*, como: "coleção de arquivos" (Bauer & Aarts, 2015, p. 54); e "coleção finita de materiais [...]" (Bauer & Aarts, 2015, p. 44) determinada de forma arbitrária pelo analista. A primeira noção é concordante com as noções da LC e da AC de linha estadunidense. Para a última noção foi destacado que os materiais devem ser homogêneos, isto é, não devem ser misturados materiais diferentes (texto e imagem, por exemplo) em um mesmo *corpus*. E esses materiais "exigem um tratamento sistemático" (Bauer & Aarts, 2015, p. 54). Os autores também sugeriram que a elaboração de um *corpus* é uma alternativa à coleta de dados, ou seja, sua constituição é funcionalmente equivalente à amostragem. A justificativa é que a amostragem não se aplica a algumas pesquisas qualitativas. Isso se contrapõe àquelas abordagens que propuseram *corpus* como uma amostra representativa de um determinado universo.

Pode-se afirmar que a maior contribuição para a concepção de *corpus* de análise é da Linguística, particularmente a LC. E é essa contribuição que determina alguns aspectos comuns entre as abordagens analíticas. Parece haver um acordo: (excetuando-se a Linguística) que o material textual não se restringe a texto, mas abrange outros materiais como imagem, áudio e vídeo; da noção de *corpus* como um conjunto; da necessidade de regras e critérios para a constituição de *corpora*. Uma diferença entre as abordagens é a natureza do conjunto (que forma um *corpus*) e as regras para sua constituição.

Metodologia

Neste trabalho investigaram-se fenômenos, compreensões e concepções, que não são objetos e não são mensuráveis. A pesquisa não se baseou no armazenamento de dados em variáveis e o posterior processamento dessas. Tampouco a generalização de resultados foi objetivo da pesquisa. Conforme disposto por Hartmut Günther (2006), houve "aceitação explícita da influência de crenças e valores sobre a teoria, sobre a escolha de tópicos de pesquisa, sobre o método e sobre a interpretação de resultados". Esses fatos são suficientes para classificar a pesquisa como qualitativa. E como será exposto adiante, embora qualitativa, esta pesquisa utilizou métodos qualitativos e quantitativos. No intuito de qualificar a pesquisa, de garantir confiabilidade e validade, buscou-se atender questões como as formuladas por Günther (2006 citado por Ollaik & Ziller, 2012), incluindo: clareza das perguntas de pesquisa; consistência do delineamento de pesquisa; explicitação de paradigmas e construtos analíticos; posição teórica e expectativas dos pesquisadores, entre outras.

A investigação de compreensões e concepções de *corpus* de análise em dissertações e teses envolveu uma análise textual com o emprego de técnica computacional, o que Neuendorf (2017, p. 146) denominou análise textual auxiliada por computador, "Computer-Aided Text Analysis" (CATA). Pelas técnicas utilizadas, a CATA se tornou zona de convergência de diferentes áreas de investigação, especialmente a Mineração de Textos (MT), a LC e a AC. Em outras palavras, a CATA aplica técnicas comuns a essas áreas de investigação.

A MT pode ser definida como "[…] um conjunto de métodos usados para navegar, organizar, achar e descobrir informação em bases textuais" (Aranha & Passos, 2006). O objetivo é extrair conhecimento oculto de grandes bases textuais e apresentá-lo de forma coerente e concisa (Bezerra & Guimarães, 2014; Faro et al., 2012; Patel & Soni, 2012). Os processos de MT possuem diferentes etapas. Entre as primeiras está o preprocessamento de texto em que as palavras, chamadas *types e tokens*, são identificadas e separadas de elementos de pontuação, de marcação e de stop words. *Types* são as palavras distintas, sem repetições, enquanto *tokens* são as palavras (types) repetidas. Stop words são artigos, pronomes, entre outras palavras consideradas irrelevantes pelo analista na análise (Patel & Soni, 2012). Em outra etapa as frequências de ocorrência de *tokens* são calculadas e as coocorrências e contextos determinados (Bezerra & Guimarães, 2014).

Similarmente à MT, a LC serve-se de técnicas como: identificação e cálculo de frequência de *tokens* — programas frequenciadores (Mello & Souza, 2012); processamento estatístico de correlações; identificação de palavras que ocorrem próximas a outra, os collocates (Sardinha, 2011); busca por palavras, expressões e padrões — programas concordanciadores (Aluísio & Almeida, 2006). Complementando, ao abordar a aplicabilidade da computação na AC, Bardin mencionou as mesmas técnicas supracitadas e afirmou que elas "[…] não são exclusivas da análise de conteúdo: a análise literária, a lexicometria, o tratamento documental da informação, por exemplo, também a praticam" (Bardin, 2011, p. 178).

Há vários programas, softwares, utilizados em CATA (Neuendorf, 2017; Piatetsky-Shapiro & Mayo, 2019), os quais se diferenciam por objetivos, resultados gerados, exigências de hardware, entre outras características. Neste trabalho utilizou-se um software livre (de código aberto), gratuito e independente de plataforma chamado Voyant Tools (Sinclair & Rockwell, 2016). O Voyant Tools é um ambiente Web de análise de *corpora* extensos, com funções de: concordanciador; frequenciador, cálculo de correlações e vínculos entre *tokens*; identificação de types; representações gráficas, entre outras. O programa possui uma lista de stop words em língua inglesa. Por esse motivo, foi necessária a introdução de um arquivo com stop words em língua portuguesa (Lopes, 2013).

Por meio do Voyant Tools conduziu-se uma análise de *corpora*. Os *corpora* foram elaborados a partir de todas as dissertações e teses disponibilizadas no sítio Web do PCM. No sítio foram listadas 203 defesas de mestrado no período de 2006 a 2018 e 65 defesas de tese no período de 2012 a 2018. Justificam-se os períodos pelas primeiras defesas de mestrado e doutorado, respectivamente, no programa, bem como o último ano com defesas registradas no sítio web. Dessas defesas estavam acessíveis os arquivos, em formato PDF, de 196 dissertações e 58 teses. Os arquivos foram convertidos em documentos de texto simples, formato ASCII ("txt"), e agrupados por ano — esse agrupamento é a unitização (Neuendorf, 2017, p. 70). Então cada *corpus* foi configurado por um arquivo, em formato ASCII ("txt"), contendo todos os documentos de um determinado ano. Geraram-se 13 *corpora* de dissertações e 7 *corpora* de teses. Os *corpora* de dissertações foram processados separadamente dos *corpora* de teses. Salienta-se que os documentos de dissertações e teses foram as fontes para a constituição dos *corpora* (os materiais analisáveis) neste trabalho.

Na investigação das dissertações (o procedimento foi repetido para as teses), todos os *corpora* foram inseridos e processados simultaneamente no Voyant Tools. Durante o processamento, efetuou-se o cálculo de ocorrência dos *types corpus* e *corpora* e suas localizações em contexto, ou *Key Word In Context* (KWIC). Estabeleceu-se para o contexto, as 30 palavras anteriores e as 30 posteriores ao termo de interesse, que foi denominado unidade de contexto (exemplo na Tabela 1). Cada unidade de contexto foi avaliada, registrando-se o que foi afirmado ou sugerido para o type (*corpus* ou *corpora*).

É importante ressaltar que todas as unidades que abordaram *corpus* como material de análise textual foram seletadas para continuidade da análise, gerando as unidades de análise. Portanto, foram descartadas as unidades de contexto que: abordavam *corpus* com o sentido de corpo (exemplos "*corpus* de conhecimento", "*corpus* teórico"); apresentaram o termo *corpus* em títulos de capítulos, subcapítulos, figuras, quadros ou tabelas; ou apenas citaram o *corpus* no processo de análise. Ressalta-se que esse foi um processo de repetidas leituras para certificação de que os itens haviam sido corretamente selecionados.

Tabela 1. Exemplo de uma unidade de contexto (KWIC)

corpus	esquerda	Termo	direita
2006	análise foram as figuras de retórica. A observação e gravação de três aulas de Ciências sobre o tema caule constituíram, juntamente com o livro didático adotado pela professor/ escola, o	*corpus*	da pesquisa e a fonte empírica para este estudo de caso (Yin, 2002) com enfoque no exame retórico dos argumentos dos livros, dos alunos e da professora (Reboul, 2004). Partimos

Fonte: elaborado pelos autores, 2020.

Fez-se então uma leitura atenta, rigorosa e criteriosa das unidades de análise, buscando as compreensões e concepções sobre *corpus*. Quando as unidades se mostraram insuficientes, recorreu-se a trechos mais amplos nos documentos originais. A leitura suscitou temas, os quais foram denominados subcategorias. Para cada subcategoria produziu-se um texto síntese, descrevendo-a e explicitando as ideias em torno de seu tema. É um processo equivalente à "produção de argumentos em torno das categorias" de Moraes (2003). Em seguida agruparam-se as subcategorias em temas mais amplos, que foram denominados categorias de análise. Nesta proposta, as categorias são resultantes da imersão do pesquisador, com suas teorias, na análise. Por isso, tratam-se de categorias emergentes, em acordo com Gamboa (2003) que ratificou a elaboração a posteriori de categorias na pesquisa qualitativa. Os textos sínteses das subcategorias propiciaram a elaboração de sínteses para as categorias, os metatextos. Os metatextos expressam concepções, compreensões, afirmações sobre *corpus* nas dissertações e teses. Moraes e Galiazzi (2011, p. 32) propuseram que os metatextos representam "um modo de teorização sobre os fenômenos investigados", sendo formados por descrição e interpretação. Segundo esses autores o nível de abstração e teorização é que determinarão um caráter mais descritivo (mais próximo do *corpus*) ou mais interpretativo (mais afastado "do material original") de um metatexto. São os metatextos que exprimem os indicadores para as compreensões e/ou concepções de pesquisadores sobre *corpus* em análises textuais.

É necessário destacar a importância do Voyant Tools ao permitir o processamento dos textos completos de dissertações e teses, não apenas fragmentos. Pode-se citar alguns exemplos da literatura que, sem o auxílio de recurso computacional, analisaram fragmentos de documentos. Santos e colaboradores (2017) relataram a análise efetuada por equipe de 10 pesquisadores sobre capítulos de metodologia em 98 dissertações de mestrado. Pedruzzi e colaboradores (2015) ao justificar a análise de 99 resumos de dissertações e teses expressaram: "Mesmo em um grupo formado por dez pesquisadores, ficaria inviável analisarmos tamanho conjunto de trabalhos. […] Avaliamos que uma média de dez resumos para cada pesquisador permitiria o desenvolvimento de uma pesquisa consistente".

Contribui para a importância do Voyant Tools a sua característica de conjugar

técnicas de diferentes correntes analíticas, tornando-a recursiva, iterativa, aberta e flexível para ser associada a outras metodologias. Nesse sentido, considera-se a análise realizada neste trabalho como exploratória. Não se trata de exploratória em um sentido superficial, mas uma análise que explora o texto a fim de revelar as informações desconhecidas.

Resultados e Discussões

Análise dos *Corpora* de Dissertações

As 196 dissertações estavam distribuídas por ano conforme apresentado na Tabela 2. No processamento desses *corpora* o Voyant Tools calculou um total de 9.087.957 de palavras (*tokens*), dos quais apenas 126.491 foram palavras distintas (types). Estimando uma média de 500 palavras por página, os *corpora*, reunidos, teriam aproximadamente 18.000 páginas. A relação entre *tokens* e *types* apontou que em média as palavras foram repetidas 71 vezes (densidade de vocabulário de 0,014). Contudo, por *corpus* essa média diminui para aproximadamente 23 repetições.

Não houve ocorrência para o termo *corpora* nos textos das dissertações e das teses. Por outro lado o termo *corpus* resultou em 292 unidades de contexto (Tabela 2) entre as dissertações. O *corpus* com maior número de ocorrências foi o de 2011 com 60 em 11 dissertações. O *corpus* com menor número de ocorrências foi o de 2007 com 2 em 12 dissertações. Esse resultado pode aguçar a curiosidade e induzir outras perguntas. Porém, a análise concentrada sobre o *corpus* de 2007, para entender o baixo número de ocorrências, não foi objetivo deste trabalho.

Tabela 2. Distribuição das dissertações e das unidades de contexto e análise por *corpus*

corpora (Conjunto de Dissertações)	**Unidades de Contexto (número de ocorrências de "corpus")**	**número de dissertações por corpus**	**Unidades de Análise (número de ocorrências selecionadas)**
2006	10	16	7
2007	2	12	1
2008	3	16	3
2009	16	21	8
2010	7	19	1
2011	60	11	39
2012	48	14	25
2013	5	12	4
2014	14	13	9
2015	15	12	4
2016	32	17	15
2017	56	16	26
2018	24	17	14
Total	**292**	**196**	**156**

Fonte: elaborado pelos autores, 2020.

Após avaliação de todas as unidades de contexto, foram selecionadas 156 (aproximadamente 53%) como unidades de análise (seleção efetuada segundo os critérios descritos na metodologia). Essas unidades de análise receberam leitura mais cuidadosa, minuciosa, aprofundada para determinação de títulos representativos, ou subcategorias. A maioria das unidades de análise foi associada a uma subcategoria, exceto 3 unidades que foram associadas a 2 subcategorias. Julgou-se não ser possível atribuir somente uma subcategoria nesses 3 casos. Obtiveram-se então 30 subcategorias (Figura 1), as quais foram avaliadas e agrupadas em temas mais amplos, as categorias. Esse agrupamento produziu 4 categorias (a posteriori): caracterização do *corpus*; definição do *corpus*; formação do *corpus*; e operação sobre o *corpus*. A distribuição das ocorrências do termo *corpus* entre as categorias foi apresentada em gráfico de setores, Figura 2a.

Subcategorias **Corpora de Dissertações**	**Subcategorias** **Corpora de Teses**	**Categorias** **resultantes**
corpus como um arquivo; *corpus* é um arquivo com "formulações linguísticas" escritas e orais; *corpus* possui "categorias (de análise)"; *corpus* possui "elementos para 'ancorar' resultados"; *corpus* possui "traços empíricos do discurso"; *corpus* possui "unidades de análise"; o *corpus* possui relações com discursos; o *corpus* está relacionado ao universo estudado;	as teorias do pesquisador determinam o *corpus*; *corpus* possui grande extensão;	caracterização do *corpus*
corpus definido como conjunto de documentos; *corpus* definido como conjunto de textos; *corpus* definido como material coletado; *corpus* definido como material contendo "algo"; *corpus* definido como material contendo informação; *corpus* definido como material contendo sentidos; *corpus* definido como material para análise; *corpus* definido como os dados da pesquisa; *corpus* definido como textos examináveis; *corpus* definido como uma ação;	*corpus* definido como conjunto de documentos; *corpus* definido como conjunto de textos; *corpus* definido como material concreto; *corpus* definido como material empírico; *corpus* definido como produções textuais; *corpus* definido como produções linguísticas sobre um fenômeno; *corpus* definido como material para análise;	definição do *corpus*

Figura 1. Subcategorias de análise para os *corpora* de dissertações e teses e categorias resultantes (continua)

Subcategorias Corpora de Dissertações	Subcategorias Corpora de Teses	Categorias resultantes
corpus formado por dados; *corpus* formado por diferentes materiais e/ou instrumentos de coleta de dados; *corpus* formado por um único tipo de material; *corpus* formado por materiais/produções textuais; indivíduos compondo o *corpus*;	*corpus* formado por dados; *corpus* formado por diferentes materiais e/ou instrumentos de coleta de dados; *corpus* formado por um único tipo de material;	formação do *corpus*
corpus fragmentado para análise; a configuração do trabalho com o *corpus*; a elaboração do *corpus* deve respeitar regras; *corpus* opcionalmente matematizado; *corpus*, material que é lido para identificação de "unidades de registro"; *corpus*, material que precisa do envolvimento dos pesquisadores; *corpus*, material utilizado para definição de categorias;	*corpus* fragmentado para análise.	operação sobre o *corpus*

Figura 1. Subcategorias de análise para os *corpora* de dissertações e teses e categorias resultantes (continuação)

Fonte: elaborado pelos autores, 2020.

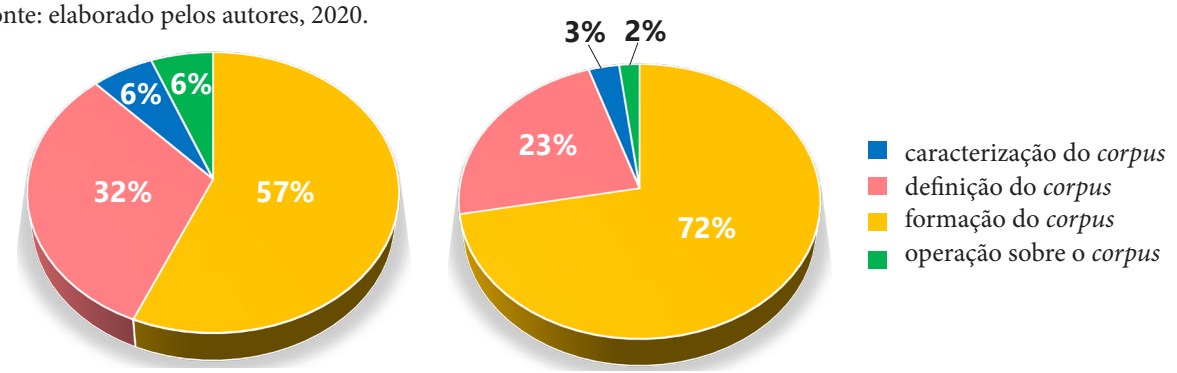

Figura 2. Porcentagem de ocorrências do termo *corpus* por categorias emergentes em: (a) dissertações; e (b) teses

Fonte: elaborado pelos autores, 2020.

No gráfico os porcentuais foram arredondados para números inteiros, por esse motivo o somatório atinge 101%. Percebe-se que a maioria das ocorrências versou sobre a formação de um *corpus*, contrastando uma minoria que abordou a caracterização de *corpus*. Esse resultado indica uma preocupação dos pesquisadores com os constituintes

de um *corpus* e talvez pouca atenção aos critérios para a constituição.

Como parte da análise, foram gerados metatextos para as categorias elencadas. Nas análises de dissertações e de teses emergiram as mesmas categorias. Por esse motivo os metatextos elaborados na análise das dissertações foram mesclados com os metatextos elaborados na análise das teses. Essas sínteses gerais estão presentes no final desta seção de resultados e discussões.

Análise dos *Corpora* de Teses

As 58 teses estavam distribuídas por ano conforme apresentado na Tabela 4. O Voyant Tools calculou um total de 4.349.636 *tokens*, 85.697 *types* e uma densidade de vocabulário de 0,02 (média de repetição de 50 vezes). Por *corpus* a média de repetição diminui para aproximadamente 21 vezes, valor comparável ao obtido para as dissertações.

Tabela 4. Distribuição das teses e das unidades de contexto e análise por *corpus*

Corpora (Conjunto de Teses)	**Unidades de Contexto (número de ocorrências de "corpus")**	**número de teses por corpus**	**Unidades de Análise (número de ocorrências selecionadas)**
2012	6	2	1
2013	41	13	17
2014	10	7	3
2015	26	11	9
2016	40	12	14
2017	20	5	5
2018	31	8	12
Total	174	58	61

Fonte: elaborado pelos autores, 2020.

O termo *corpus* ocorreu em 174 unidades de contexto (Tabela 4) entre as teses. O *corpus* com maior número de ocorrências foi o de 2016 com 40 em 12 teses. A avaliação das unidades de contexto gerou 61 unidades de análise (35% das unidades de contexto). Na sequência da análise obtiveram-se 13 subcategorias. Também registraram-se 3 unidades de análise associadas a 2 subcategorias. Das 13 subcategorias para os *corpora* de teses, apenas 5 foram coincidentes com aquelas para os *corpora* de dissertações: *corpus* definido como material para análise; *corpus* formado por dados; *corpus* formado por diferentes materiais e/ou instrumentos de coleta de dados; *corpus* formado por um único tipo de material; *corpus* fragmentado para análise. Porém, essas subcategorias foram organizadas nas mesmas 4 categorias definidas anteriormente. A distribuição das ocorrências do termo *corpus* entre as categorias foi apresentada na Fig. 2b. O gráfico da Fig. 2b apresenta o mesmo perfil do gráfico da Fig 2a, confirmando a indicação de uma preocupação com constituintes de *corpus* e pouca atenção aos critérios de constituição.

Metatextos das Categorias

Neste trabalho foram elaborados metatextos de caráter mais descritivo. Contudo, após a apresentação de cada metatexto, foi incluído um exercício interpretativo com "intuições e entendimentos atingidos a partir da impregnação intensa […]" (Moraes & Galiazzi, 2011, p. 37) com os *corpora* analisados. Conforme mencionado previamente, os metatextos na análise dos *corpora* de dissertações foram elaborados independentemente dos metatextos na análise dos *corpora* de teses e só mesclados ao final.

A primeira categoria resgatada para a discussão é Caracterização do *Corpus*. Esta é a categoria em que as unidades de análise apresentaram alguma caracterização do *corpus*. Duas propostas analíticas destacaram-se como referências entre os trabalhos envolvidos pelas ocorrências: a AD de linha francesa (Moreira, 2012; Orlandi, 2015) e a ATD de Moraes e Galiazzi (2011). Uma primeira característica é que um *corpus* reflete as teorias que orientam e acompanham a pesquisa do analista. O *corpus* é elaborado e organizado em acordo com as orientações dessas teorias. Entre as unidades de análise identificou-se também o *corpus* com característica de um arquivo com materialidade e composto por "formulações linguísticas", na forma escrita e/ou oral, produzidas por sujeitos de uma pesquisa. O *corpus* possui um "dizer" que é constituído pela relação com outros discursos, bem como possui "traços empíricos do discurso" de um indivíduo ou grupo social. Esses traços são identificados pelo analista de discurso quando reorganiza o *corpus*. O *corpus* também possui elementos, que extraídos, permitem ao pesquisador "ancorar" resultados em uma realidade empírica. Do *corpus* são obtidas unidades de análise, as quais acompanham categorias de análise. E as categorias são obtidas da imersão do pesquisador sobre o material de análise e do agrupamento de elementos de significação semelhantes. Além de unidades de análise uma categoria é acompanhada por um metatexto. Outra característica para o *corpus* é sua relação com o universo estudado, isto é, "a constituição do *corpus* diz respeito ao universo estudado em sua totalidade e a formulação e reformulação de hipóteses e objetivos". Uma última característica, identificada em alguns trabalhos que utilizaram a ATD, é que um *corpus* possui grande extensão e por isso o analista trabalha com parte do *corpus*.

Moraes e Galiazzi (2011, p. 39) sugeriram que "os metatextos não devem ser entendidos como modos de expressar algo já existente nos textos, mas como construções do pesquisador com intenso envolvimento de sua parte". Assim, com base na sugestão desses autores efetuou-se um exercício de reflexão sobre o metatexto elaborado. Inicia-se a reflexão com a ideia de *corpus* como arquivo. O *corpus* como arquivo é um espaço em que se mantém um conjunto de documentos ou materiais. Um arquivo pode ser "aberto", permitindo que sejam inseridos, removidos ou acessados os documentos contidos nele e considerando-se as tecnologias de informação digitais, pode-se propor um *corpus* como arquivo eletrônico. Por outro lado, Aiub (2012), citando Pêcheux, colocou que "arquivo é '[…] entendido, no sentido amplo, de campo de documentos pertinentes e disponíveis sobre uma questão' […] o arquivo é […] organizado por uma leitura"; o *corpus* é construído a partir da leitura de arquivos. O *corpus* como material textual é elaborado a partir de

um código linguístico. Caso seja considerada a possibilidade de um material textual de origem visual, deve-se considerar que esse material seja transposto do objeto visual para um com os sinais e estrutura de um código linguístico. A compreensão da mensagem deve depender do reconhecimento desse código e sua adequada interpretação. As mensagens fazem parte dos discursos registrados em um *corpus*. Além das mensagens, um *corpus* apresenta elementos de pensamento, de conhecimento, de concepções, de crenças, de história, entre outros, daqueles que emitiram um ou mais discursos. Esses elementos é que devem permitir que os resultados da análise textual sejam conectados à "realidade empírica" (o contexto, o espaço e o período daquilo que é investigado). Isso é uma forma de garantir confiabilidade e validade da análise (Moraes, 2003). Para o analista, explorar um *corpus* passa por teorizações, por leituras imersivas e exaustivas, por formulação de perguntas e hipóteses, por fragmentação do material textual, entre outros. E é na fragmentação que o analista pode selecionar unidades relacionadas a assuntos ou temáticas, as quais podem ser agrupadas em temas mais amplos denominados posteriormente de categorias (categorias que emergem do *corpus* a partir do processo de análise). Nas unidades de análise não foi encontrada determinação, ou mensuração, do que seja um *corpus* de "grande extensão".

A segunda categoria resgatada é Definição do *Corpus*. A categoria reúne as unidades de análise que apresentaram uma definição para *corpus*. Percebeu-se que as compreensões de *corpus* entre as dissertações foram influenciadas predominantemente por duas propostas analíticas: a AC de Bardin (2011) e a ATD. A proposta de AD destacou-se entre as teses. Nas unidades de análise, de um modo geral e independente da proposta analítica, o *corpus* é um material produzido ou selecionado por um pesquisador para sua pesquisa. Adicionalmente *corpus* foi definido como produções textuais, um conjunto de documentos ou conjunto de textos. As produções textuais também são reconhecidas como produções linguísticas que abordam, ou referem-se, a um ou mais fenômenos de um determinado contexto e período. Um *corpus* é lido, examinado e analisado, "por impregnação intensa", segundo critérios estabelecidos pelo pesquisador em um processo que envolve procedimentos e técnicas analíticas. Em alguns trabalhos baseados na ATD destacaram-se três compreensões para *corpus*: "conjunto de documentos que representa as informações da pesquisa" (e as informações podem auxiliar na elaboração de categorias de análise); material do qual são extraídos sentidos no processo de análise; material sobre o qual a análise textual se concretiza. Nas unidades de análise em que *corpus* foi definido apenas como material coletado para ser estudado na pesquisa, não há identificação de seu tipo (texto, áudio, imagem, entre outros). Quando definido o tipo texto, o *corpus* foi compreendido como o conjunto de dados da pesquisa, tendo surgido inclusive a expressão "*corpus* de dados". E o conjunto de documentos contendo dados é selecionado de forma a garantir resultados que sejam confiáveis. Dependendo da proposta metodológica de análise (ATD, AD) o *corpus* será reconhecido como "material concreto" ou empírico, isto é, o material contendo os dados obtidos ou coletados de instrumentos, registros, entre outras fontes. Em uma unidade

de análise encontrou-se uma definição totalmente divergente da ideia de *corpus* como um material. Nela o *corpus* consistiu na aplicação de uma oficina, ou seja, uma ação.

No metatexto a impregnação intensa pode ser entendida como o processo em que o analista se debruça concentrada e intensamente sobre o *corpus*, intuindo, inferindo, deduzindo, identificando categorias, entre outros. Outra ideia destacada diz respeito aos sentidos extraídos de um *corpus*. Esses podem ser as ideias que não estão explícitas, mas que são identificadas através das interpretações, intuições e deduções do analista. Por último, a compreensão de *corpus* como conjunto que representa as informações de pesquisa, oriunda de uma dissertação de 2012 que usou ATD, se contrapõe a uma citação a Guilhaumou e Maldidier (1997) em uma das dissertações de 2011, que um arquivo "não pode ser traduzido como um documento do qual se retiram informações".

A terceira categoria é Formação do *Corpus*. Esta categoria reúne as unidades de análise que abordaram como um *corpus* é formado. Na maioria dessas unidades verifica-se que o *corpus* é formado por um ou mais materiais, com ou sem especificação destes. Em três unidades de análise surgiu uma indicação divergente e incomum: *corpus* composto por indivíduos (exemplo "[…] dos bolsistas constituintes do *corpus* de pesquisa [...]"). Houve indicação de *corpus* como matéria-prima da pesquisa e constituído de produções textuais (em referência a ATD), bem como indicação de *corpus* formado por dados (não definido como conjunto de dados, nem formado exclusivamente por esses). Destaca-se em uma unidade de análise a menção "*corpus* de dados". Em outra unidade a afirmação que os textos que compõem o *corpus* são os dados propriamente ditos e que "[...] todo dado torna-se informação a partir de uma teoria, [...] 'nada é realmente dado', mas tudo é construído". A maioria das unidades de análise aponta para *corpus* formado por um ou por diferentes tipos de materiais. Em algumas situações não há diferenciação entre os instrumentos de coleta de dados e os dados coletados através destes.

Dentre os materiais constituintes de *corpus* exemplificam-se: livros didáticos, capítulos de livros didáticos, registros de observação, aulas, gravação de aulas, atas de eventos, registros em jornais, filmagens, documentos históricos, atividade prática (a atividade ou um registro da atividade?), entrevistas, questionários, transcrições de entrevistas, transcrições de questionários, diários de campos, depoimentos, levantamento bibliográfico, referências bibliográficas, análise documental, transcrições de aulas, excertos de discussões, análise epistemológica, avaliação formal, matrizes de avaliação, textos de apoio, planejamentos didáticos, áudios de cursos, referências bibliográficas, referenciais teóricos, relatos de experiência, relatos, artigos, revistas, produções didático-pedagógicas, capítulos de livros didáticos, documentos, dissertações e tese (dissertações e tese foram considerados um mesmo tipo de material), narrativas, diários, escritos de Leonardo da Vinci, cartazes (análise de imagens), atividade experimental, atividades de uma sequência didática. Em algumas unidades de análise os autores descreveram um *corpus* formado por um único tipo de material, por exemplo *corpus* constituído por transcrições de entrevistas. Em outras unidades de análise os autores descreveram um *corpus* formado por mais de um tipo (diferente) de material, por exemplo "o

corpus constitui todos os dados coletados: as respostas dos questionários, as narrativas produzidas e as transcrições das entrevistas realizadas com os estudantes, além dos documentos do curso, como o PPP […]".

Uma primeira reflexão sobre o metatexto desta categoria é que dados constituem o material que é processado para gerar informação. Se o analista extrai do *corpus* excertos para sua análise, significa que esses são os dados, não o *corpus* integralmente. Nessa situação, os dados não abrangem, necessariamente, a totalidade dos materiais textuais que compõem um *corpus*. Uma segunda reflexão trata do *corpus* formado por diferentes tipos de materiais, indicando a falta de homogeneidade e de distinção entre *corpus* e *corpora*. Isso pode justificar, parcialmente, a não ocorrência do type *corpora* nas dissertações e teses. A última reflexão trata de *corpus* constituído por indivíduos. As três unidades de análise que apresentaram a compreensão de *corpus* formado por indivíduos foram extraídas de duas dissertações de mestrado, uma de 2016 e outra de 2018, envolvendo pesquisadores de um mesmo grupo de pesquisa. Nas duas dissertações foram propostos *corpora* formados por indivíduos (bolsistas de Programa Institucional de Bolsa de Iniciação à Docência) e por materiais textuais (diários e respostas de questionários, respectivamente) – *corpus* formado por diferentes materiais. Adicionalmente, a dissertação de 2018 é a origem da definição de *corpus* como aplicação de oficina, comentada na categoria Definição de *Corpus*. É preciso lembrar que todos os referenciais tratam de materiais textuais como constituintes de *corpus*, pois a análise é textual. E em acordo com a categoria Operação sobre o *Corpus*, na análise efetuam-se operações sobre um *corpus* (o *corpus* não é uma operação/ação). Um indivíduo não é um material textual, nem é submetido a uma operação. Contudo ele pode ser fonte de materiais textuais quando externaliza algo por diferentes linguagens. Tratar um indivíduo como componente de um *corpus* é incompatível com os referenciais e denota incompreensão do conceito de *corpus* de análise.

A última categoria formada foi Operação sobre o *Corpus*. Esta categoria reúne unidades de análise que mencionam algum tipo de operação sobre o *corpus*. Tratam-se de operações para a elaboração do *corpus* ou de análise do *corpus*. Para a elaboração do *corpus* devem ser seguidas regras, as quais agregarão confiabilidade e validade à análise. Nesse caso foram citadas algumas das regras que constam na obra Análise de Conteúdo de Bardin (2011): regra da exaustividade, regra da amostragem, regra da homogeneidade, regra da pertinência. Na análise o analista envolve-se com o *corpus*, o lê intensivamente para compreendê-lo e identificar unidades de registro. De forma similar, em uma tese foi mencionada a necessidade de fragmentação do *corpus* no processo de análise, para permitir o destaque de trechos importantes denominados unidades constituintes. Essa fragmentação é efetuada a partir de critérios estabelecidos pelo analista, em acordo com teorias e metodologias por ele escolhidas. No processo de leitura intensiva o analista também separa excertos (recortes), e identifica categorias e subcategorias. Uma unidade de análise, citando Bardin (2011), colocou que "A categorização é uma operação de classificação de elementos constitutivos de um conjunto [...]". Esse envolvimento com o

corpus ocorre em função da relação entre a teoria e a análise (uma referência a proposta de AD), das características dos dados, dos objetivos do analista e das questões por ele formuladas. Em uma unidade de análise foi destacada que a aplicação de procedimentos matemáticos no processo de análise é uma opção do analista, na qual se supera a mera quantificação e busca-se a compreensão de regularidades do discurso no *corpus*.

Do metatexto pode-se colocar que em uma pesquisa existem ações que são efetuadas sobre o *corpus*, as quais podem ocorrer no processo de elaboração ou de análise do *corpus*. A leitura, por exemplo, é uma ação que não produz um efeito sobre um *corpus*, não o altera. Uma fragmentação, por outro lado, é uma operação sobre o *corpus*. O envolvimento do analista com o *corpus* implica um processo intenso, metódico, recursivo de leitura, de compreensão, de apreensão, de interpretação, entre outros. E nesse envolvimento com o *corpus* o analista pode combinar procedimentos quantitativos e qualitativos, mantendo a característica qualitativa da análise e desviando-se do dualismo mencionado por Gamboa (2003). Nesse sentido, a unidade de análise que destacou os procedimentos matemáticos é importante por indicar uma dissertação que empregou exame estatístico e AD, exemplificando assim a superação do dualismo pesquisa qualitativa e quantitativa.

Em uma avaliação global, confirmaram-se três referenciais majoritariamente utilizados pelos pesquisadores do PCM: a obra de Laurence Bardin; as obras de Eni Orlandi; a obra de Roque Moraes e Maria do C. Galiazzi. As menções sobre *corpus* de análise nos documentos são quase reproduções das obras citadas, sem uma reflexão sobre as discussões promovidas pelos autores dos referenciais. A impressão provocada é que houve, de modo geral, uma aplicação sem uma apropriação epistemológica das abordagens analíticas. Ao mesmo tempo em que as concepções convergem ao reproduzir os mesmos referenciais, elas divergem por uma desarmonia com esses referenciais. A divergência foi flagrante em concepções como: *corpus* composto por indivíduos; *corpus* como aplicação de uma oficina; *corpus* formado por instrumentos de coleta de dados (ao invés dos dados coletados com esses instrumentos).

Entre alguns dos referenciais utilizados neste trabalho (Bardin, 2011; Bauer & Aarts, 2015; Orlandi, 2015; Sánchez, 1995) há o apontamento de necessidade de critérios a serem obedecidos na construção de um *corpus*. Retornando às categorias obtidas, verifica-se que não houve a apresentação de critérios detalhados, claros e específicos (em acordo com a metodologia analítica adotada) utilizados na elaboração dos *corpora*. Apenas registrou-se uma menção à necessidade de regras, particularmente aquelas propostas por Bardin (2011, p. 126) que têm caráter mais amplo. A incerteza em relação a critérios é intensificada pelo exposto no metatexto da categoria formação do *corpus*: *corpus* constituído por diferentes tipos de materiais. Isso conflita com o critério apresentado por Bauer e Aarts (2015) de homogeneidade de materiais em um *corpus*. Materiais diferentes são organizados, preparados e tratados de forma distinta. Por exemplo, a análise de uma imagem não segue o mesmo protocolo da análise de uma entrevista em áudio (Bauer & Gaskell, 2015).

De fato não se verificou nas unidades de análise o uso, ou resgate, de outros referenciais teóricos, como a LC, ou especialmente a linha estadunidense de AC. Tais referenciais aprimorariam concepções, compreensões sobre as metodologias adotadas e cuidados na constituição de *corpus*. Seriam evitadas, por exemplo, afirmações de *corpus* formado por diferentes materiais, ou confusões entre dados e informações.

Conclusões e Implicações

Pode-se colocar que, em relação ao universo investigado, as menções de *corpus* (para uma análise textual) estão vinculadas às metodologias analíticas adotadas, majoritariamente a AC (francesa), a AD (francesa) e a ATD. Apesar deste vínculo, foram evidenciadas fragilidades nas compreensões e concepções do que deve ser um *corpus* e de como ele deve ser elaborado e processado. De um modo geral, *corpus* foi considerado um mero conjunto, algo reunido para uma pesquisa. Essas fragilidades estão relacionadas a dois afastamentos. O primeiro é o afastamento das concepções sobre *corpus* em relação aos referenciais adotados. Os referenciais foram citados, reproduzidos, mas não foram apropriados adequadamente. O segundo é o afastamento entre os referenciais adotados e destes em relação à Linguística. Não se identificou uma caracterização de *corpus* comum e independente das metodologias e referenciais utilizados. E a noção de *corpus* originária da Linguística parece ter sido perdida. Esses apontamentos foram, de certo modo, corroborados pelas bancas avaliadoras das dissertações e teses. Isso não significa omissão, mas provavelmente uma naturalização do uso do termo *corpus* sem um aprofundamento conceitual.

Os resultados obtidos determinam algumas implicações para as pesquisas na área de Educação em Ciências Naturais. Por exemplo, há um indicativo nos resultados que os pesquisadores atentaram pouco a critérios para elaboração de *corpora*. Entretanto, um *corpus* válido (e que confira validade à análise) precisa ser elaborado em acordo com critérios bem definidos. Mesmo que as metodologias analíticas apresentem concepções de *corpus* distintas, eles são necessários. Aliás, se compreende possível estabelecer alguns critérios comuns a tais metodologias. Um exemplo é a garantia de disponibilidade e de acessibilidade a outros pesquisadores. Sugere-se nesse caso a geração de arquivos eletrônicos, seguindo padrões para formato, armazenamento e transferência. Outro exemplo está relacionado à homogeneidade de um *corpus*, isto é, constituição por um único tipo de material. Julga-se importante que materiais diferentes sejam sistematizados em *corpora* separados. Em razão da necessidade de critérios, pretende-se em trabalho subsequente, apresentar um conjunto para a elaboração de *corpora* que poderá ser aplicado em várias metodologias analíticas na pesquisa em Educação em Ciências. Ressalta-se que esse conjunto de critérios poderá ratificar o *corpus* como elemento que estabelece aproximação entre as diferentes metodologias de análise.

Um *corpus* pode ser considerado uma fonte de dados sistematizados em uma análise textual, independentemente se todo o seu conteúdo ou apenas excertos serão analisados. A análise textual, propriamente dita, é iniciada com um *corpus* e efetuada

sobre este. Em princípio os dados representam um certo fenômeno em um determinado contexto e período, por isso esses dados permitem ao analista conectar resultados de uma análise a uma realidade empírica. Isso é importante, mas é insuficiente para garantir confiabilidade e validade à análise. O *corpus* como fonte de dados precisa estar disponível e acessível a outros pesquisadores, o que se configura em outro requisito para confiabilidade e validade. Defende-se, a partir dos aspectos apresentados e de referenciais como Bauer e Gaskell (2015), que se considere o *corpus* como um critério para a confiabilidade e validade de uma análise textual. Se um pesquisador elege um conjunto de materiais textuais para analisá-los, tornar-se-á fundamental a constituição de um *corpus* ou *corpora*.

Outro assunto importante é a polissemia do termo *corpus*. Surgiram nas unidades de contexto expressões que não se relacionam à análise textual, por exemplo: "*corpus* do conhecimento físico", "*corpus* geográfico", "*corpus* de saberes", "*corpus* teórico", "*corpus* filosófico", entre outras. Ou seja, alguns documentos utilizaram o termo *corpus* com mais de um sentido. Para minimizar esse problema, sugere-se a expressão "*corpus* de análise" quando a referência for ao material elaborado para uma análise textual. Complementando a sugestão, seria interessante também que pesquisadores em Educação em Ciências consultassem referenciais da Linguística, especialmente a LC, para melhor compreender a noção de *corpus* de análise.

Ampliando as considerações sobre *corpus* de análise, salienta-se que neste trabalho foram produzidos *corpora* de análise extensos. O uso de recurso computacional foi essencial para a investigação com tais *corpora*. Em realidade, além de identificar as compreensões sobre *corpus*, procurou-se mostrar a viabilidade de um processo analítico que concilia CATA, análise categorial e elaboração de metatextos. É um processo que tem potencial para ampliar os horizontes de pesquisa qualitativa em Educação em Ciências. A primeira razão é a possibilidade de analisar *corpora* de análise extensos e não apenas documentos fragmentados ou limitados. A segunda razão é que um pesquisador pode combinar diferentes programas, qualitativos e/ou quantitativos, para robustecer o processo de análise. As limitações estarão primordialmente relacionadas às capacidades de processamento e armazenamento. Enquanto as dificuldades estarão relacionadas ao nível de engajamento e de capacitação de um pesquisador para usar os recursos computacionais de análise textual.

Complementando, alguns resultados gerados pelo programa Voyant Tools instigam análises que não foram escopo deste trabalho (exemplo, o porquê apenas duas ocorrências do termo *corpus* entre as dissertações de 2007). Tais análises podem ser objeto de investigação em outras propostas de pesquisa, quer estejam em desenvolvimento, quer sejam futuras. Para isso, os dados que suportam os resultados deste estudo, arquivos de dissertações e teses, foram derivados dos seguintes recursos disponíveis em domínio público: vínculo Produções no sítio web do PCM; <http://www.pcm.uem.br/dissertacoes>; <http://www.pcm.uem.br/teses>.

Por fim, os resultados apresentados não são generalizáveis. Mas em acordo ao

que foi mencionado inicialmente neste trabalho, é possível predizer que em um universo mais abrangente de pesquisadores da área de Educação em Ciências Naturais, deverão ser verificadas as mesmas compreensões e concepções. Isto exige desses pesquisadores uma autoavaliação e revisão de fundamentos teóricos e metodológicos que orientam suas pesquisas.

Referências

Aiub, G. F. (2012). Arquivo em Análise do Discurso: Uma breve discussão sobre a trajetória teórico-metodológica do analista. *Leitura, 2*(50), 61–82. https://doi.org/10/ggxxm3

Aluísio, S. M., & Almeida, G. M. de B. (2006). O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio, 4*(3), 156–178. https://doi.org/10.4013/6002

Aranha, C., & Passos, E. (2006). A Tecnologia de Mineração de Textos. *Revista Eletrônica de Sistemas de Informação, 5*(2), 1–8. https://doi.org/10/ggmkf7

Bardin, L. (2011). *Análise de Conteúdo* (L. A. Reto & A. Pinheiro, Trads.; 1a ed). Edições 70.

Bauer, M. W., & Aarts, B. (2015). A construção do *corpus*: Um princípio para a coleta de dados qualitativos. In M. W. Bauer & G. Gaskell (Orgs.), *Pesquisa qualitativa com texto, imagem e som: Um manual prático* (13a ed, p. 39–63). Vozes.

Bauer, M. W., & Gaskell, G. (Orgs.). (2015). *Pesquisa qualitativa com texto, imagem e som: Um manual prático* (13a ed). Vozes.

Bezerra, C. A., & Guimarães, A. J. R. (2014). Mineração de texto aplicada às publicações científicas sobre gestão do conhecimento no período de 2003 a 2012. *Perspectivas em Ciência da Informação, 19*(2), 131–146. https://doi.org/10/ggmk7q

Caregnato, R. C. A., & Mutti, R. (2006). Pesquisa qualitativa: Análise de discurso versus análise de conteúdo. *Texto & Contexto - Enfermagem, 15*(4), 679–684. https://doi.org/10/dkczmh

Carlomagno, M. C., & Rocha, L. C. da. (2016). Como Criar e Classificar Categorias para Fazer Análise de Conteúdo: Uma Questão Metodológica. *Revista Eletrônica de Ciência Política, 7*(1), 173–188. https://doi.org/10/gd7gbz

Chrysostomo, T. da S., & Messeder, J. C. (2017). Uso Da Publicidade Televisiva na Sala de Aula: Percepções e Contribuições de Acadêmicos de Licenciatura em Química. *Revista Areté | Revista Amazônica de Ensino de Ciências, 10*(22), 281–293. http://periodicos.uea.edu.br/index.php/arete/article/view/650

Faro, A., Giordano, D., & Spampinato, C. (2012). Combining literature text mining with microarray data: Advances for system biology modeling. *Briefings in Bioinformatics, 13*(1), 61–82. https://doi.org/10/dq79ks

Gamboa, S. A. S. (2003). Pesquisa Qualitativa: Superando tecnicismos e falsos dualismos. *Revista Contrapontos, 3*(3), 393–405. https://siaiap32.univali.br/seer/index.php/rc/article/view/735

Glosbe. (2020). *Corpus* em Português — Latim-Português Dicionário. In *Glosbe dicionário [em linha]*. https://pt.glosbe.com/la/pt/*corpus*

Günther, H. (2006). Pesquisa qualitativa versus pesquisa quantitativa: Esta é a questão? *Psicologia: Teoria e Pesquisa, 22*(2), 201–209. https://doi.org/10/db5743

Informática, P. (2013). Definição de *corpus* no Dicionário Priberam da Língua Portuguesa, o dicionário online de português contemporâneo. In *Dicionário Priberam da Língua Portuguesa [em linha]*. https://dicionario.priberam.org/*corpus*

Lopes, A. (2013). Portuguese stop words. GitHub Gist. https://gist.github.com/alopes/5358189

Mello, H. R. de, & Souza, R. R. (2012). A linguagem da ciência: Prospecção de dados baseados em *corpora*. STIS Seminários Teóricos Interdisciplinares do SEMIOTEC - *Cadernos Didáticos e Anais, 1*(1), 19p. http://www.periodicos.letras.ufmg.br/index.php/stis/article/view/2115

Moraes, R. (2003). Uma tempestade de luz: A compreensão possibilitada pela análise textual discursiva. *Ciência & Educação (Bauru), 9*(2), 191–211. https://doi.org/10/dv5vc4

Moraes, R., & Galiazzi, M. do C. (2011). *Análise Textual Discursiva* (2a ed). Editora Unijuí.

Moreira, L. A. L. (2012). Análise do Discurso no Brasil: Reflexões acerca de sua construção teórico-metodológica. *Leitura, 2*(50), 109–133. https://doi.org/10/ggxxts

Neuendorf, K. A. (2017). *The content analysis guidebook* (Second edition). SAGE.

Oliveira, E., Ens, R. T., Andrade, D. B. S. F., & Muss, C. R. (2003). Análise de Conteúdo e Pesquisa na Área da Educação. *Revista Diálogo Educacional, 4*(9), 11. https://doi.org/10/ggkxdw

Ollaik, L. G., & Ziller, H. M. (2012). Concepções de validade em pesquisas qualitativas. *Educação e Pesquisa, 38*(1), 229–242. https://doi.org/10/gfwgvp

Orlandi, E. P. (2015). *Análise de discurso: Princípios & procedimentos* (12a ed). Pontes.

Patel, F. N., & Soni, N. R. (2012). Text mining: A Brief survey. *International Journal of Advanced Computer Research*, 2(4), 234–239. https://pdfs.semanticscholar.org/11c4/6d00a0e136e8e4e27aa15fbb8c9111cdee75.pdf

Pedruzzi, A. das N., Schmidt, E. B., Galiazzi, M. do C., & Podewils, T. L. (2015). Análise Textual Discursiva: Os movimentos da metodologia de pesquisa. *Atos de Pesquisa em Educação, 10*(2), 584–604. https://doi.org/10/ggkxv2

Piatetsky-Shapiro, G., & Mayo, M. (2019). *Text Analysis, Text Mining, and Information Retrieval Software*. KDnuggets. https://www.kdnuggets.com/software-for-data-mining-analytics-data-science-and-knowledge-discover/text-analysis-text-mining-and-information-retrieval-software/

Pinhão, F., & Martins, I. (2009). A Análise do Discurso e a Pesquisa em Ensino de Ciências no Brasil: Um Levantamento da Produção em Periódicos entre 1998 e 2008. *Anais do VII Encontro Nacional de Pesquisa em Educação em Ciências*, 12. http://posgrad.fae.ufmg.br/posgrad/viienpec/pdfs/518.pdf

Rocha, D., & Deusdará, B. (2005). Análise de Conteúdo e Análise do Discurso: Aproximações e afastamentos na (re)construção de uma trajetória. *Alea: Estudos Neolatinos,* 7(2), 305–322. https://doi.org/10/bs2cvm

Sánchez, A. (1995). *Cumbre: Corpus lingüístico del español contemporáneo: fundamentos, metodología y aplicaciones*. Sociedad General Española de Librería.

Santos, A. R. dos, Sousa, R. S. de, & Galiazzi, M. do C. (2018). A Análise Textual Discursiva na Pesquisa em Educação Química: A Categorização como Possibilidade de Ampliação de Horizontes. *Iniciação & Formação Docente, 4*(2), 167–178. http://seer.uftm.edu.br/revistaeletronica/index.php/revistagedeles/article/view/2250

Santos, B. F., Vaz, Á. S., Leite, P. L., Barbosa, C. S., Araújo, L., Lyra, A. B., Santos, B., Moreira, C. B., Santana, M. L. A. D., & Martins, R. B. (2017). O estudo dos métodos de análise em dissertações como aprendizagem e formação de pesquisadores para a pesquisa qualitativa: Relato de uma experiência. *Revista Brasileira de Pós-Graduação, 14*, 1–17. https://doi.org/10/ggkzb3

Sardinha, T. B. (2000). Lingüística de *Corpus*: Histórico e problemática. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 16*(2), 323–367. https://doi.org/10/ct75cj

Sardinha, T. B. (2011). Metáforas e Linguística de *Corpus*: Metodologia de análise aplicada a um gênero de negócios. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 27*(1), 01–20. https://doi.org/10/ggmk7x

Sinclair, S., & Rockwell, G. (2016). *Voyant Tools*. Voyant Tools. https://voyant-tools.org/

Julio Murilo Trevas dos Santos

https://orcid.org/0000-0002-5691-9265
Universidade Estadual de Maringá
Programa de Pós-Graduação em Educação para a Ciência e a Matemática
Maringá, Paraná, Brasil
Universidade Federal da Fronteira Sul
Campus Realeza
Realeza, Paraná, Brasil
jtrevas@uffs.edu.br

Neide Maria Michelan Kiouranis

https://orcid.org/0000-0002-1279-9994
Universidade Estadual de Maringá
Departamento de Química
Maringá, Paraná, Brasil
nmmkiouranis@gmail.com

Submetido em 03 de março de 2020

Aceito em 17 de julho de 2020

Publicado em 19 de agosto de 2020