

UM ESTUDO QUANTITATIVO DO TERMO *LINGUÍSTICA* EM PROJETOS PEDAGÓGICOS DE CURSOS DE LETRAS-PORTUGUÊS DE UNIVERSIDADES FEDERAIS

D. R. Sales¹

Orientador: Prof. Dr. M. G. C. Martins²

RESUMO

O presente trabalho tem o objetivo de analisar como o termo *linguística* é referido em projetos pedagógicos de cursos (PPC) de Letras-Português de universidades públicas federais. Justifica-se esta pesquisa com base nas Diretrizes Curriculares Nacionais, nomeadamente o Parecer CNE/CSE de nº 492, de 03 de abril de 2001 (BRASIL, 2001), e a Resolução CNE/CES de nº 18, de 13 de março de 2002 (BRASIL, 2002), que orientam que, durante a formação do aluno da licenciatura em Letras, deve haver um eixo vinculado à área de Linguística, que contemple, ao mesmo tempo, uma visão crítica das teorias adotadas nesse campo de conhecimento e uma visão analítica da estrutura e do funcionamento da língua estudada. Para realizar esta pesquisa, faz-se uma análise quantitativa textual de PPCs dessas licenciaturas, em particular de títulos, ementas e indicações bibliográficas de todas as disciplinas desses PPCs, publicamente disponibilizados nas páginas eletrônicas de instituições federais de ensino superior. A extração de dados foi realizada com o suporte aos programas Orange (DEMŠAR, Janez et al.) e Sketch Engine (KILGARRIFF et al., 2014). Utilizei como critério de inclusão no corpus, os PPC de habilitação única em Letras-Português, de cursos em atividade, de instituições públicas federais, na modalidade presencial e com nota 4 ou 5 (de acordo com o índice de CPC). Desse modo, foi possível identificar quais universidades se assemelham com relação à seleção do léxico em torno do item *linguística*.

Palavras-chave: análise quantitativa de textos; colocação; frequência; agrupamento; linguística.

ABSTRACT

The present work aims to analyze how the term *linguistics* is referred to in pedagogical projects of Portuguese Language courses (PPC) at federal public universities. This research is justified based on the National Curriculum Guidelines, namely Opinion CNE/CSE No. 492, of April 3, 2001 (BRASIL, 2001), and Resolution CNE/CES No. 18, of March 13, 2002 (BRASIL, 2002), which guide that, during the formation of the student of the degree in Languages, there must be an axis linked to the area of Linguistics, which contemplates, at the same time, a critical view of the theories adopted in this field of knowledge and an analytical view of the structure and functioning of the studied language. To carry out this research, a quantitative textual analysis of the PPCs of these degrees was carried out, in particular the titles, menus and bibliographical indications of all the disciplines of these PPCs, publicly available on the electronic pages of federal institutions of higher education. Data extraction was performed with support for Orange (DEMŠAR, Janez et al.) and Sketch Engine (KILGARRIFF et al., 2014) programs. I used as a criterion for inclusion in the corpus, the PPC of single qualification in Portuguese-Language, of courses in activity, of federal public institutions, in the face-to-face modality and with grade 4 or 5 (according to the CPC index). In this way, it was possible to identify which universities are similar in relation to the selection of the lexicon around the linguistic item.

Keywords: quantitative analysis of texts; placing; frequency; clustering; linguistics.

¹ Graduanda do curso de Letras Português da Universidade Federal Rural do Semi Árido (UFERSA)

² Docente da Universidade Federal Rural do Semi-Árido (UFERSA).

1. INTRODUÇÃO

As Diretrizes Curriculares Nacionais (DCN), nomeadamente o Parecer CNE/CES de nº 492, de 03 de abril de 2001 (BRASIL, 2001) e a Resolução CNE/CES de nº 18, de 13 de março de 2002 (BRASIL, 2002), orientam que o graduado em Letras deverá apresentar múltiplas competências e habilidades adquiridas durante sua formação acadêmica convencional, teórica e prática, e, também, fora dela. Esses documentos apontam que, no percurso formativo do aluno da licenciatura em Letras, deve haver um eixo de formação vinculado à área de Linguística, que contemple, ao mesmo tempo, uma visão crítica das teorias adotadas nesse campo de conhecimento e uma visão analítica da estrutura e do funcionamento da língua estudada.

O interesse por essa pesquisa se justifica pela necessidade de o professor de Letras-Português ter domínio do conhecimento científico que se produz na área da linguística, que se faz necessária para a sua atuar como professor. Quanto a isso, contamos com um documento oficial para nortear a formação docente, a Base Nacional Comum para Formação de Professores da Educação Básica (BNC-Formação) que apresenta como sua primeira competência, “Compreender e utilizar os conhecimentos historicamente construídos para poder ensinar a realidade com engajamento na aprendizagem do estudante e na sua própria aprendizagem colaborando para a construção de uma sociedade livre, justa, democrática e inclusiva.” (BRASIL, 2019, p. 13) Assim, não estudamos as teorias linguísticas apenas para decorar ou aprender aquelas teorias de forma superficial, e sim com o objetivo de aprender os seus fundamentos e conseguir pensar em como esses conhecimentos devem ser repassados para os alunos, de modo que venham a agregar ainda mais nas suas formações.

A Linguística no decorrer dos mais de cem anos de existência construiu muitos conhecimentos ao longo desse tempo, com diferentes perspectivas teóricas sobre os fenômenos e criando diferentes teorias que abordam formas diferentes de compreender a linguagem enquanto fenômeno da língua, de modo que foram criadas diferentes escolas para estudar as mais variadas teorias linguísticas que surgiram com o passar do tempo e as suas evoluções. Os desdobramentos linguísticos que mais se destacam com o passar do tempo, são: o estruturalismo, gerativismo e o funcionalismo. Foi no ano de 1916 que ocorreu o ponto de partida que deu início aos estudos da linguística como ciência, com a publicação do Curso de Linguística Geral, do estudioso Ferdinand de Saussure.

Esse estudioso referia-se à Linguística na sua abordagem estrutural, chamando a de sistema e em seguida começando a ser referida como estrutura, essa teoria estuda o sistema da língua, observando o modo como eles são organizados e compõem outras unidades, estudando

e descrevendo como as diversas línguas existentes se organizam. Em seguida, Noam Chomsky aborda uma nova teoria que contraria a estudada anteriormente, intitulada de gerativismo. Assim, ele buscou analisar um modelo teórico formal, com o intuito de descrever e analisar a linguagem humana de acordo com a sua faculdade mental funcional. Posteriormente, surgiu a abordagem funcionalista, que tem Halliday como o seu principal estudioso. Diferentemente das demais abordagens, essa aborda a questão usual da língua, estudando a língua como um todo, estudando as funções realizadas pelos elementos linguísticos.

Tendo a Linguística produzido tantos conhecimentos e havendo hoje uma necessidade de o professor ter competência de saber esses conteúdos e ensinar, como a *linguística* está sendo proposta nos PPCs dos cursos de Letras-Português das universidades federais. Logo, é importante identificar as teorias linguísticas que são abordadas na formação desses licenciandos, como informam os documentos norteadores dos cursos de formações profissionais, as Diretrizes Curriculares Nacionais (DCN), nomeadamente o Parecer CNE/CES de nº 492, de 03 de abril de 2001 (BRASIL, 2001), ao orientar que

O objetivo do Curso de Letras é formar profissionais interculturalmente competentes, capazes de lidar, de forma crítica, com as linguagens, especialmente a verbal, nos contextos oral e escrito, e conscientes de sua inserção na sociedade e das relações com o outro. [...] deve ter domínio do uso da língua ou das línguas que sejam objeto de seus estudos, em termos de sua estrutura, funcionamento e manifestações culturais, além de ter consciência das variedades linguísticas e culturais. Deve ser capaz de refletir teoricamente sobre a linguagem, de fazer uso de novas tecnologias e de compreender sua formação profissional como processo contínuo, autônomo e permanente. [...] O profissional deve, ainda, ter capacidade de reflexão crítica sobre temas e questões relativas aos conhecimentos linguísticos e literários. (BRASIL, 2001, p. 30)

Portanto, os cursos de letras apresentam como objetivos de que os seus formandos devem ter domínio das teorias linguísticas e as suas diversas modalidades que são apresentadas no decorrer do curso, para que possam refletir sobre como se relacionam com o ensino e buscar a melhor forma de ensinar a língua aos educandos, à luz dessas teorias.

Para responder à pergunta de investigação que estabeleço, apresento aqui uma análise documental de projetos pedagógicos de cursos (PPC) de licenciaturas em Letras-Português de universidades públicas federais brasileiras. A escolha por PPCs se justifica no fato de que é nesse documento - institucionalizado, público e acessível online - que as convicções teóricas de um dado curso de graduação estão devidamente explicitadas. O PPC, como descrevem Seixas, Coelho-Lima, Silva e Yamamoto, materializa essas convicções teóricas à luz de um cenário

global, em que tais convicções estão em relação dinâmica com diferentes camadas da realidade humana no processo de formação profissional. Nas palavras de Seixas e colegas, o PPC é

um documento normativo dos cursos de graduação que apresenta características de projeto com informações acerca da concepção e da estruturado curso e seus elementos reguladores internos. Nos PPCs estão presentes aspectos técnicos normativos, concepções de homem e de sociedade, além de um componente político fundamental, sendo elemento agregador de diversas instâncias da realidade, desde sua dimensão cotidiana dos cursos até diretrizes das políticas macroeconômicas. (SEIXAS; COELHO-LIMA; SILVA; YAMAMOTO, 2013, p. 114)

Para a identificação dos modos como a linguística é apresentada nos PPCs, opto por realizar uma análise de colocados, com base em frequências e agrupamentos, com o termo linguística nos PPCs. A colocação é uma técnica de análise consolidada na linguística de *corpus*, em que se entende, segundo EVERT (2008, p. 1212-1213), que colocações são “combinações de palavras características e frequentemente recorrentes”, nas quais “o significado e o uso de uma palavra (o nó) podem, até certo ponto, ser caracterizados por suas colocações mais típicas”. Colocações, portanto, são “co-ocorrências recorrentes restritas lexicalmente e/ou pragmaticamente de pelo menos dois itens lexicais que estão em uma relação sintática direta entre si”³ (BARTSCH, 2004, *apud* EVERT, 2008, p. 1213). Objetiva-se, neste artigo, apresentar uma análise dos colocados com o item linguística (e variações) a fim de identificar como se apresenta esse item nos PPCs de licenciaturas em Letras. Noutras palavras, tem-se o intuito de identificar quais as principais abordagens teóricas a que os licenciandos são expostos no decorrer da sua formação.

Deste modo, o presente texto está organizado a princípio com a introdução, seguida de uma seção intitulada de enquadramento teórico, ao qual apresenta ao leitor os princípios base de uma análise quantitativa automatizada de texto e alguns critérios que são abordados no decorrer da pesquisa. Além disso, nesse mesmo tópico são apresentados os principais domínios linguísticos estudados. No próximo ponto são abordados os procedimentos metodológicos que foram utilizados no decorrer da pesquisa para atingir os objetivos propostos. A última parte apresenta os resultados e discussões encontradas no decorrer da pesquisa, seguida das considerações finais.

3 Minha tradução de: “lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other”

2. ENQUADRAMENTO TEÓRICO

Nesta seção, apresento princípios teóricos que orientam esta pesquisa. Começo por discutir os pressupostos de uma pesquisa quantitativa (automatizada) de textos. A análise quantitativa automatizada de texto, neste estudo, é assumida como um método. No entanto, há um conjunto de princípios teóricos que justificam discuti-la nesta seção. De seguida, apresento correntes teóricas da Linguística que julgo mais expressivas para o entendimento de um quadro panorâmico no processo de formação inicial docente.

2.1. Análise quantitativa de texto

A análise quantitativa automatizada de texto utiliza o léxico como o seu principal objeto de estudo, permitindo “a aplicação de métodos para se extrair [desse léxico] inferências estatísticas de populações de textos” (ROBERTS, 2000, p. 259). Dutt-Ross e Cruz (2021, p. 192) referem que a análise quantitativa de textos é “[...] uma técnica que envolve a coleta de dados qualitativos em uma plataforma digital de conteúdo público para realizar uma análise quantitativa destes conteúdos por meio da apresentação de análise de correlação entre os termos utilizados pelos usuários.” A análise quantitativa de textos é um método de trabalho vinculado às Ciências Sociais e à Linguística, nomeadamente à linguística de *corpus* e computacional, muitoem parte porque tal método de análise pode ser realizado em textos de mídia social, como feeds do Twitter, postagens do Facebook, blogs e dados transcritos de sites de compartilhamento de vídeos, como o Youtube e outros.

Esse tipo de estudo consolida a perspectiva de que uma língua apresenta, em seu uso – nos textos, como entendo –, regularidades estatísticas, uma das razões por que George Zipf se tornou famoso. Mais especificamente, a Lei de Zipf, como lembra Cantos Gómez (2002, p. 235), define que há uma relação constante entre a classificação de uma palavra em uma lista de frequências e a frequência com que ela é usada em um texto. Essa propriedade estatística de uma língua natural, ainda em linha com Cantos Gómez, revela-se em padrões linguísticos, que são regulares e independentes de um falante, escritor ou assunto, e revela também que o comportamento linguístico de uma dada comunidade está de acordo com algumas expectativas: padrões quantitativos ou estatísticos. Isso se ilustra, por exemplo, no fato de que as palavras mais frequentes numa língua são de menor extensão do que as palavras mais incomuns, tipicamente mais longas. Os dados quantitativos, porém, precisam ser interpretados, com interpretações que condigam com os resultados numéricos, que os tornem coerentes, por isso umapesquisa quantitativa também pode ser de cunho qualitativo, interpretativista.

Os dados de uma análise quantitativa são exibidos como frequências e coocorrências (colocações), em que uma determinada palavra de interesse – como aqui, “linguística” – pode ter destaque no processo de análise. Também, os resultados de extração de dados podem ser exibidos com base em agrupamentos, que sistematizam modos de organização de documentos a partir das ocorrências do seu léxico. Esses procedimentos são explicados a seguir.

2.1.1. Frequência

O conceito de frequência é um dos fundamentos de muitos procedimentos metodológicos da linguística de corpus, como se vê em cálculos de dispersão ou de colocação. Frequências podem ser apresentadas em números brutos ou em números relativos (proporção ou percentual), o que permite comparar diferentes corpora ou diferentes fatos linguísticos (lexicais ou gramaticais) dentro do mesmo corpus. É importante enfatizar que, as frequências não são explicadas, no entanto, é necessário analisar as concordâncias para entender e as frequências dos termos que sobressaem, como lembram Baker et al. (2006, p. 76),

as frequências não se explicam: análises baseadas em concordância são, portanto, necessárias para explicar por que certas palavras são mais frequentes do que outras. Também pode ser necessário levar em conta as frequências de termos relacionados (...).

2.1.2. Colocação

A colocação se constitui pelo fato de que certas palavras são mais prováveis de ocorrer em combinação com outras palavras em certos contextos. Uma colocação é, portanto, uma palavra que ocorre na vizinhança de outra palavra. Há diferentes programas informáticos em que se pode calcular frequências de colocação, com a especificação da janela de ocorrências e por diferentes modelos estatísticos, como informação mútua, o Z-score, MI3, log-log ou escores de probabilidade logarítmica (BAKER et al., 2006). As colocações podem ser úteis para demonstrar a existência de viés ou conotação nas palavras.

2.1.3. Agrupamentos (cluster)

Uma análise de agrupamentos é uma técnica estatística adotada pela linguística de corpus que permite a identificação de categorias automaticamente, calculando graus de semelhança ou diferença entre vários textos (BAKER et al., 2006). Como define Doni (2004, p. 10), os componentes agrupados refletem "padrões consistentes de comportamento", de forma que cada agrupamento "seja similar de acordo com algum critério ou métrica". A partir dos agrupamentos, é possível estabelecer subconjuntos de análise mais compreensíveis.

2.2. A Linguística como ciência

A Linguística é a ciência responsável por estudar a linguagem humana e os estudos que analisam os mais variados processos de mudanças que a língua sofre no decorrer dos tempos, apresentando inúmeras conceituações, dependendo das diversas teorias linguísticas que estão em uso no momento, variando de acordo com os conceitos que os estudiosos abordam em cada um deles.

A publicação do Curso de Linguística Geral, de Ferdinand Saussure (1916), no início do século XX, foi o momento precursor no que diz respeito aos estudos linguísticos, uma vez que, é o momento em que a Linguística passa a ser considerada como uma nova ciência, como destaca Fiorin (2010), ao definir o objetivo da ciência linguística como aquela que “[...] estuda a principal modalidade dos sistemas sócio-sígnicos, as línguas naturais, que são a forma de comunicação mais altamente desenvolvida e de maior uso.” A Linguística passa a ser considerada como a ciência responsável por estudar e descrever a linguagem humana e as suas diversas formas, no primeiro momento ela estudava apenas a linguagem verbal, oral ou escrita, com o passar dos tempos e o surgimento de novas teorias, a linguística passa a estudar a linguagem como um todo e os seus vários tipos.

Com a evolução da Linguística, conseqüentemente, começam a aparecer correntes teóricas de abordagens distintas. Dentre elas, as que mais se destacam são: o estruturalismo, gerativismo e funcionalismo, e ao falar nessas correntes linguísticas os nomes que se destacam, consecutivamente são Saussure, Chomsky e Halliday. Dentre esses estudos, a primeira grande teoria que surgiu foi o Estruturalismo, que tem Saussure como o seu principal estudioso e a autora Orlandi (1992) aborda essa corrente teórica da seguinte forma:

Essa organização interna da língua, que Saussure chama sistema, seus sucessores chamarão estrutura. Com essa noção, procuram valorizar a ideia de que cada elemento da língua só adquire um valor quando se relaciona com o todo de que faz parte. [...] O método que analisa a língua assim definida é o método estrutural e dá à linguística a posição de ciência-piloto das ciências humanas. (ORLANDI, 1992, p. 23-24)

Assim, temos o estruturalismo como a teoria pioneira no que diz respeito aos estudos linguísticos, desse modo, servindo como base para a criação dos novos estudos que surgiram posteriormente, que possuem maneiras distintas de compreender o fenômeno da linguagem. Como é o caso do gerativismo, que surgiu na tentativa de se opor aos estudos estruturalistas, utilizando de objetos de estudos que não eram utilizados pela corrente anterior, com isso surgiu a necessidade da criação de uma nova abordagem, que recebeu este nome.

O estudioso Noam Chomsky se destaca ao tratar dessa teoria, seu nome é referência ao falar sobre os estudos linguísticos, uma vez que ele contribuiu grandemente com as suas teorias e com isso, os estudos linguísticos sucessores ao estruturalismo costumam utilizar das suas teorias linguísticas como referência das correntes que deve seguir.

Marcuschi (2016, p. 19) define o gerativismo como um aparato teórico que “é rigoroso e busca dar conta de forma ordenada, explicativa, econômica e teoricamente adequada de fenômenos abstratos e universais da língua.” O Gerativismo é uma nova forma de visualizar a Linguística, contribuindo para a criação de novos objetos de estudos da língua, apontando uma nova forma de observar os aspectos linguísticos.

Como uma contrapartida às teorias linguísticas anteriores, surgiu o funcionalismo, na tentativa de se opor aos demais estudos, como também de acrescentar novas teorias e posicionamentos das teorias linguísticas. Sperança-Crisuolo (2014) afirma que nessa teoria se “encontram critérios para investigar a língua em sua totalidade, desde a produção de um enunciado até sua compreensão” (SPERANÇA-CRISCUOLO, 2014, p. 8). Portanto, essa teoria dedica-se mais ao estudo do uso ao invés de focar apenas no estudo formal da língua, tendo Halliday como o principal nome ao falar sobre os estudos funcionalistas.

A Linguística é uma ciência diversificada e apresenta variações com relação aos seus objetos de estudos, que variam de acordo com as diferentes correntes existentes. A Psicolinguística é um dos objetos de estudos da ciência linguística, ela estuda a relação predominante entre a língua e os seus processos psicológicos, considerando o modo como é adquirido a aquisição da linguagem, como reage com isso, a sua evolução e o seu processamento na mente humana. Sempre considerando essa relação da linguagem com a mente e a sua aprendizagem. Já a Sociolinguística é uma corrente linguística que visa levar em consideração a língua como um fator social, considerando a existência de uma relação entre a linguagem e a sociedade. Assim, considera esses fatores como sendo indissociáveis e isso pode ser considerado como a base na relação existente entre o homem e a sociedade. Como destacam Mussalin e Bentes (2009, p. 31):

[...] o objeto da Sociolinguística é o estudo da língua falada, observada, descrita e analisada em seu contexto social, isto é, em situações reais de uso. Seu ponto de partida é a comunidade linguística, um conjunto de pessoas que interagem verbalmente e que compartilham um conjunto de normas com respeito aos usos linguísticos.

Com o passar dos tempos, os estudiosos começaram a sentir a necessidade de criar uma área de linguística que realizasse análises das teorias baseadas no uso real da língua, desse modo criaram a Linguística de Corpus, que é a área da Linguística responsável por realizar uma coletade dados relacionados a uma dada temática que se deseja estudar e, com isso, cria-se um corpus

de análise para que os estudos venham a ser desenvolvidos com base nele. Assim, os estudos sobre os casos de usos da língua tornam-se mais fáceis e simples de ser realizados.

[...] a linguística de corpus não define somente uma metodologia emergente para o estudo da linguagem, mas uma nova maneira de fazer pesquisa, e de fato uma nova abordagem filosófica para este assunto. O computador, como uma ferramenta tecnológica de poder indiscutível, tornou este novo tipo de linguística possível [...] (LEECH, 1992, *apud* SVARTVIK, 1996, p. 12)

Então, entendemos nessa seção que a Linguística é considerada como uma ciência responsável por estudar a linguagem humana e as suas diferentes teorias. Uma vez que é uma ciência que vem evoluindo no decorrer dos mais de cem anos de existência, criando diversas ramificações das três principais correntes, assim, surgindo novas correntes linguísticas e objetos de estudos, fazendo com que esses estudos evoluam a cada dia mais. Sabendo disso, na seção seguinte serão apresentados os procedimentos metodológicos utilizados para o cumprimento dos objetivos propostos desta pesquisa, apresentando as etapas realizadas para a elaboração dessa análise.

3. PROCEDIMENTOS METODOLÓGICOS

Nesta seção, apresento a descrição dos métodos de trabalho escolhidos para cumprir com o objetivo desta pesquisa. Para tanto, pauto-me, sobretudo, em Gil (2022).

Esta pesquisa caracteriza-se como uma pesquisa de cunho descritivo, pois que, sendo o seu objetivo geral analisar como a Linguística é referida em projetos pedagógicos de cursos de Letras-Português de universidades públicas federais, há clara intenção de "identificar características de determinada população ou fenômeno" (GIL, 2022, p. 42). Aqui, o que se quer identificar, como já referido noutras partes deste artigo, é o modo como, no momento presente, o conjunto de teorias do campo linguístico, acumulado ao longo de mais de um século de história, se materializa institucionalmente nos processos de formação inicial de professores de Língua Portuguesa. Para uma consideração eficiente do objetivo, entram em jogo as seguintes variáveis: quanto ao curso, considera-se a nota CPC (Conceito Preliminar de Curso), e quanto à IES, a sua localização geográfica. Essas variáveis podem permitir uma caracterização mais qualificada dos resultados, que serão extraídos de uma análise quantitativa da variável textual, que se constitui dos títulos e ementas de todas as disciplinas dos PPCs selecionados.

Como nesta pesquisa se extraem dados dos PPCs de cursos de Letras-Português, pelo que são "documentos institucionais, mantidos em arquivos de (...) órgãos públicos" (GIL, 2022, p. 44) e como se tem "o propósito de fornecer respostas a um problema (...) específico" (GIL, 2022, p. 74), que aqui é saber como se adota a ciência linguística nos cursos de Letras-Português

das universidades federais, esta pesquisa se delineia como uma pesquisa documental, tendo sido respeitadas as etapas que asseguram a realização de uma pesquisa dessa natureza, a saber: a) formulação do problema e dos objetivos; b) identificação das fontes; c) localização das fontes e acesso aos documentos; d) avaliação dos documentos; e) seleção e organização das informações; f) análise e interpretação dos dados; g) redação do relatório (GIL, 2022, p. 74). A seguir, descrevem-se essas etapas situadas nesta pesquisa, exceto a primeira, já apresentada na introdução deste artigo.

Constituem-se como fontes desta pesquisa os projetos pedagógicos de cursos de Letras-Português de universidades federais disponibilizados nas páginas eletrônicas dos cursos, de acesso livre (sem necessidade de *login*, de cadastro ou de autorização de uso). Esses documentos tipicamente estão disponibilizados em formato PDF e se encontram em um único arquivo. É justamente a localização dos arquivos, em páginas eletrônicas oficiais das IES, que garante a autenticidade e a credibilidade dos documentos aqui utilizados, dois dos critérios de qualidade, como afirma Gil (2022, p. 76).

Além dos critérios acima referidos, deve-se considerar a representatividade e o significado para a avaliação da qualidade dos documentos. Para efeitos de representatividade do corpus de pesquisa, verificou-se que, a partir de dados exportados do e-MEC⁴ - pela conjugação de nome do curso e rótulo CINE⁵, há, no Brasil, 69 cursos de licenciatura presencial em Letras-Português em atividade nas universidades federais, o que se configura como universo desta pesquisa. Não estão incluídos neste número os cursos de dupla habilitação e os cursos de Português enquanto língua estrangeira. Como recorte de intervenção no corpus, estabeleceu-se que seriam analisados apenas os PPCs de licenciaturas cuja nota CPC fosse 4 ou 5, obtida na última avaliação e como disponibilizado no site do e-MEC. A nota CPC (conceito preliminar de curso), como se descreve no site do Ministério da Educação, é

o conceito que avalia o curso, em uma escala de 1 a 5. Para o cálculo, são considerados: Conceito Enade (desempenho dos estudantes na prova do Enade); Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD); corpo docente (informações do Censo Superior sobre o percentual de mestres, doutores e

⁴ Disponível em: <https://emec.mec.gov.br/>. Acesso em 30 de outubro de 2022.

⁵ Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais (Cine Brasil) (cf. <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/cine-brasil/classificacao>). Acesso em 30 de outubro de 2022.

regime de trabalho) e percepção dos estudantes sobre seu processo formativo (informações do Questionário do Estudante do Enade). (BRASIL, 2022)

Portanto, assume-se a nota CPC como o indicador mais global da qualidade do curso.

Outro critério assumido para a constituição do corpus foi a exclusão de PPCs de mesma universidade, prevalecendo como parte do escopo de análise o PPC de curso criado mais recentemente. Tendo sido aplicado tal critério, obteve-se um total de 26 documentos para a constituição do corpus de análise, o que representa aproximadamente 38% do universo de pesquisa. No entanto, por questões técnicas, como, por exemplo, a inoperância da página eletrônica da universidade no momento da coleta, não foi possível obter os PPCs de onze dessas instituições, razão por que o corpus desta pesquisa conta com 15 PPCs, que representam 22% do universo estudado.

Assim, compõem o corpus desta pesquisa os PPCs das seguintes universidades:

- Universidade Federal do Ceará (UFC)
- Universidade Federal Rural do Rio de Janeiro (UFRRJ)
- Universidade Federal de Santa Catarina (UFSC)
- Universidade Federal do Mato Grosso (UFMT)
- Universidade Federal do Tocantins (UFT)
- Universidade Federal de São João Del-Rei (UFSJ)
- Universidade Federal do Acre (UFAC)
- Universidade Federal do Maranhão (UFAM)
- Universidade Federal de Pelotas (UFPEL)
- Universidade Federal do Rio Grande do Norte (UFRN)
- Universidade Federal da Paraíba (UFPB)
- Universidade Federal do Rio Grande (FURG)
- Universidade Federal do Pampa (UNIPAMPA)
- Universidade Federal de Goiás (UFG)
- Universidade Federal de São Paulo (UNIFESP)

O último critério de avaliação da qualidade dos documentos, como lembra Gil (2022, p. 76), é o seu significado, o que implica considerar "se o conteúdo do documento está situado dentro de seu contexto histórico, em relação a outros documentos que o cercam e o que significa, e também ao que significa em termos de pesquisa". Esse critério se garante pelo próprio caráter institucional dos PPCs, que devem ser continuamente atualizados, sendo um dos

critérios de avaliação de cursos de graduação definidos no Sistema Nacional de Avaliação da Educação Superior (SINAES), conforme Brasil (2004).

Como assevera Gil (2022, p. 76), numa análise documental, é necessário que os dados obtidos sejam reduzidos ao mais relevante. No contexto desta pesquisa, optou-se por centrar a análise nas seções textuais dos PPCs em que o termo linguística pudesse ser apresentado em contextos mais específicos, permitindo a extração das informações que se desejava apresentar nesta pesquisa. Desse modo, constituem matéria de análise desta pesquisa apenas os títulos das disciplinas com respectivas ementas e os títulos das obras que compõem as referências bibliográficas dessas disciplinas.

Esta pesquisa se propõe a uma análise interpretativista de dados obtidos a partir de uma análise quantitativa automatizada do léxico do corpus. Uma análise quantitativa se sustenta no princípio de que "os sentidos das trocas comunicativas podem ser encontrados por meio de análises matemáticas sobre as palavras, que passam a ser tratadas estatisticamente por meio da contagem de frequência e testes multivariados (...)" (OLIVEIRA, 2019, p. 11). Aqui, emprego nomeadamente os procedimentos de identificação de similaridades por agrupamento dos documentos (*cluster analysis*) e a frequência e a coocorrência de termos, que se detalham tanto na extração de n-gramas, como na construção de redes de colocados.

Para o tratamento computacional, realizou-se uma etapa de limpeza do corpus para a sua inserção nas plataformas digitais de análise. Especificamente criou-se um arquivo no formato CSV em editor de texto, com a indicação da universidade, da nota e do texto enquanto cabeçalhos. Por não serem interessantes a esta pesquisa, realizou-se a exclusão de elementos gráficos não lexicais, tais como tabulações, ícones de listas e enumerações, logotipos etc.

O corpus de análise resultou com um total de 25.788 *tokens* e 3.836 *types*, distribuídos por 15 documentos. A quantidade de *tokens* por documento para constituição do corpus se descreve detalhadamente na **Erro! Fonte de referência não encontrada.**, abaixo:

Tabela 1: Distribuição da quantidade de *tokens* por documento na constituição do corpus

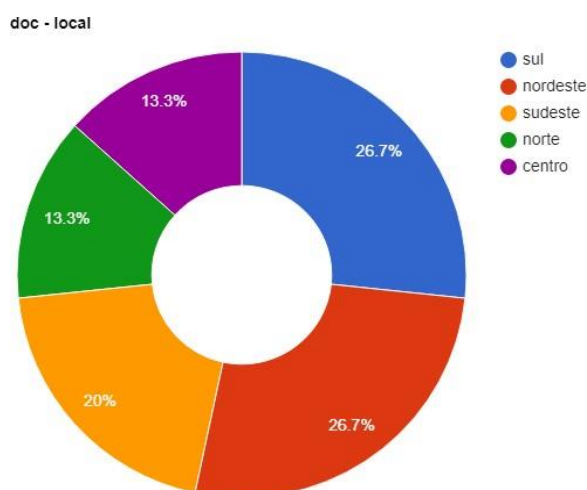
Documento	Quantidade de <i>tokens</i> por documento (em números absolutos)	Quantidade de <i>tokens</i> por documento em relação ao corpus (em números percentuais)
uft.txt	5140	10,95
ufsc.txt	4808	10,24
furg.txt	3697	7,87
ufsj.txt	3689	7,86
ufpel.txt	3315	7,06
ufpb.txt	3254	6,93
ufac.txt	3218	6,85

ufrn.txt	3107	6,62
ufmt.txt	3094	6,59
unifesp.txt	3091	6,58
ufma.txt	2535	5,40
ufc.txt	2290	4,88
unipampa.txt	2235	4,76
ufg.txt	1894	4,03
ufrj.txt	1586	3,38

Como se pode ver, os documentos contribuem com diferentes quantidades de *tokens* na constituição do corpus, em que UFT e UFSC cobrem mais de 21% do corpus total e UFRRJ representa apenas 3,38% do total. Essas diferenças se diluem na aplicação dos modelos estatísticos usados pelos softwares de tratamento de textos.

Abaixo se apresenta a distribuição da localização geográfica das IES:

Gráfico 1 - Distribuição de documentos no corpus por localização geográfica.



Fonte: Autoria própria

O Gráfico 1 revela que, entre as universidades com nota 4 ou 5 que compõem o corpus, há predominância de universidades nordestinas e sulistas (4 de cada região), seguidas de universidades do sudeste (3), com menores ocorrências de universidades do norte e do centro-oeste (2 de cada região).

Destaca-se, por fim, que das universidades selecionadas, apenas uma recebeu CPC 5 na última avaliação (UFSJ), representando 6,7% do total de documentos do corpus.

A extração de dados desse corpus e a sua relevância para o objetivo desta pesquisa se apresentam na seção seguinte, em que também os discuto.

Para fins de comparação, apresenta-se a rede de colocados com o item *linguística*, da categoria nome, e com os vínculos da categoria adjetival. Formando as classes de textos que apresentam uma maior aproximação, uma vez que fazem uso de palavras semelhantes que são distribuídas estatisticamente de modo igual, estando essas palavras sempre próximas e assim agrupam-se de acordo com as suas similaridades. Então formando esses agrupamentos de classes, para melhor identificar os textos semelhantes.

Figura 1: Rede de colocados com o item linguística no corpus total



Sobre os agrupamentos (C1, C2 e C3), os outros procedimentos de mineração de dados foram aplicados, nomeadamente a obtenção de frequências, de n-gramas e a obtenção de rede de colocados, objetivo último desta pesquisa.

4.1. Cluster 1

Quanto ao cluster 1, de que fazem parte dados textuais dos PPCs da UFC e da UFRRJ, primeiramente extraíram-se as 20 palavras (lemas) mais frequentes, apresentadas abaixo comparativamente às dez palavras (lemas) mais frequentes no corpus total:

Tabela 2: Dez palavras (lemas) mais frequentes no Cluster 1, em comparação ao corpus total

CORPUS				C1			
Posição	Item	Valores absolutos	Valores percentuais	Posição	Item	Valores absolutos	Valores percentuais
1	português	775	165,06	1	português	110	2,84

2	literatura	720	153,34	2	literatura	92	2,37
3	língua	519	110,54	3	estudo	69	1,78
4	ensino	402	85,62	4	língua	56	1,44
5	estudo	359	76,46	5	linguístico	32	0,83
6	brasileiro	334	71,14	6	brasileiro	32	0,83
7	introdução	274	58,36	7	teoria	27	0,70
8	linguístico	273	58,14	8	análise	26	0,67
9	análise	273	58,14	9	história	22	0,57
10	texto	271	57,72	10	ensino	22	0,57

As palavras que aparecem com maior frequência no cluster 1 apresentam uma grande semelhança com as palavras que frequentemente aparecem no corpus total, o quantitativo de 8 das 10 palavras que são elencadas no C1 também são recorrentes no corpus total. A diferença que encontramos é de acordo com as ordens de recorrências, já que a palavra “português” é a mais frequente, seguida de “literatura” que aparecem nas mesmas colocações como as principais palavras apresentadas nesses textos, as demais aparecem em frequências diferentes.

O lema que mais aparece no C1 é “português”, com os valores absolutos de 110, isso ocorre por conta dos PPCs em análise fazerem parte dos cursos de Letras-português e a aparição dessa palavra ser recorrente em várias partes dos textos, como é o caso de um grande número das ementas das disciplinas, como também as disciplinas na sua maioria apresentam o nome português no seu título, como por exemplo, “Português brasileiro”; “Estilística do Português”; “Tópicos de Português”; “Nova gramática do português contemporâneo.” e várias outras palavras. Esse lema não mantém relação apenas com o termo “linguística”, ele também mantém aproximação com as teorias literárias e consequentemente os termos relacionados.

O termo “linguística” não aparece com tantas recorrências no corpus e apresenta uma maior colocação no C1, no entanto os termos relacionados a “literatura” ainda se sobressaem, com isso, entende-se que nestes documentos que fazem parte do cluster 1, a menção das teorias literárias é mais recorrente.

O resultado da extração de bigramas – com, no mínimo, duas ocorrências – mais recorrentes no C1, que, em termos sintéticos, são dois itens que estão contíguas no corpus, vê-se na tabela abaixo, com valores absolutos e com valores percentuais, que considera o total de ocorrências de bigramas:

Tabela 3: Dez bigramas (tokens) mais frequentes no Cluster 1, em comparação ao corpus total

CORPUS	C1
--------	----

Posição	Bigrama	Valores absolutos	Valores percentuais	Posição	Bigrama	Valores absolutos	Valores percentuais
1	língua portuguesa	259	0,55	1	língua portuguesa	44	1,14
2	literatura portuguesa	131	0,28	2	literatura portuguesa	24	0,62
3	literatura brasileira	130	0,28	3	literatura brasileira	21	0,54
4	ensino fundamental	75	0,17	4	literaturas africanas	5	0,13
5	prática pedagógica	58	0,13	5	portuguesa estudo	4	0,10
6	estudos literários	49	0,09	6	literatura cearense	4	0,10
7	língua latina	41	0,08	7	literatura infantil	4	0,10
8	português brasileiro	40	0,08	8	gramática latina	4	0,10
9	ensino médio	34	0,07	9	unidade curricular	3	0,08
10	educação básica	34	0,07	10	principais teorias	3	0,08

Os bigramas apresentados com base nos lemas mais frequentes nesses textos, confirmamos que foi falado acima, que as teorias literárias se sobressaem, tanto na C1 como no corpus como um todo, pois os nomes mais recorrentes são “língua portuguesa”, que pode estar relacionado tanto com as teorias linguísticas como literárias, seguida por “literatura portuguesa” correspondendo as disciplinas do viés literário, depois “literatura brasileira”. Assim, concluo que de modo geral, no cluster 1 as teorias e estudos literários se destacam.

Quando se consideram apenas os bigramas em que conste o item “linguística”, identificou-se na amostra geral e no subcorpus C1 o que se dispõe na tabela abaixo:

Tabela 4: Dez bigramas (tokens) com o item *linguística* mais frequentes no Cluster 1, em comparação ao corpus total

CORPUS				C1			
Posição	Bigrama	Valores absolutos	Valores percentuais	Posição	Bigrama	Valores absolutos	Valores percentuais
1	linguística geral	27	0,06	1	linguística estudo	2	0,05
2	linguística textual	21	0,04	2	linguística cognitiva	2	0,05
3	linguística aplicada	19	0,04	3	análise linguística	2	0,05
4	análise linguística	18	0,04				
5	linguística descritiva	8	0,02				

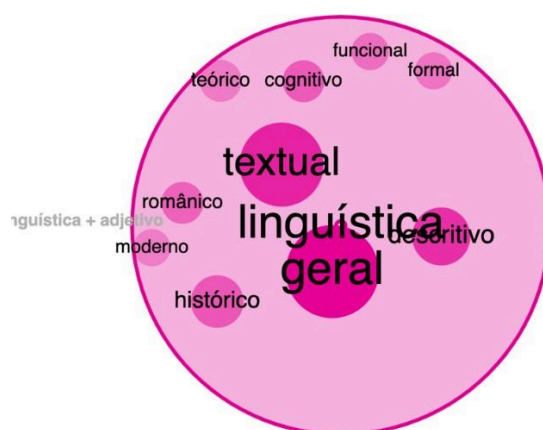
6	linguística histórica	7	0,01				
7	mudança linguística	7	0,01				
8	variação linguística	6	0,01				
9	linguística teórica	3	0,01				
10	linguística cognitiva	3	0,01				

No que diz respeito aos bigramas que aparecem com o termo “linguística”, como é apresentado acima na tabela 4, é possível identificar quais recorrências do termo linguística prevalece no corpus e no C1. O corpus apresenta “linguística geral” com o bigrama de maior recorrência, sendo 27 o seu valor absoluto, isso se dá pelo grande uso do livro de Saussure como base dos estudos linguísticos, assim a maioria das instituições de ensino utilizam como referência bibliográfica de grande parte das suas disciplinas.

Os bigramas do termo “linguística” do cluster 1 comparados ao corpus, mostra uma menor frequência dos bigramas relacionados ao termo pesquisado, mostrando uma recorrência percentual de apenas 0,05 das palavras total do texto, e esses bigramas não apresentam frequências de termos relacionados as teorias linguísticas de maiores destaques nos bigramas do corpus. Com isso, certificamos os baixos índices de termos ligados a linguística presentes nos textos do C1.

Por fim, a extração de colocados, também centrada no nome “linguística”, e associada a termos adjetivos revelou o seguinte cenário de conexões cluster 1:

Figura 2: Rede de colocados com o item linguística no C1



A rede de colocados da figura 1, exemplifica o que foi mostrado acima, ao falar que o termo *linguística* é mais frequente seguida do termo *geral*, no que diz respeito ao livro de Ferdinand Saussure, *Curso de Linguística Geral*. Nessa rede de colocados os lemas com os círculos maiores são os termos mais recorrentes nos textos e os menores são os que aparecem em menor frequência.

4.2. Cluster 2

Quanto ao cluster 2, de que fazem parte dados textuais dos PPCs da UFSC, UFMT, UFT, UFSJ, UFAC, UFMA e UFPB, a extração das 20 palavras (lemas) mais frequentes se detalha abaixo, também em comparação com as dez palavras mais frequentes no corpus total:

Tabela 5: Dez palavras (lemas) mais frequentes no Cluster 2, em comparação ao corpus total

CORPUS				C2			
Posição	Item	Valores absolutos	Valores percentuais	Posição	Item	Valores absolutos	Valores percentuais
1	português	775	165,06	1	português	341	1,32
2	literatura	720	153,34	2	literatura	321	1,25
3	língua	519	110,54	3	língua	261	1,01
4	ensino	402	85,62	4	ensino	217	0,84
5	estudo	359	76,46	5	estudo	168	0,65
6	brasileiro	334	71,14	6	análise	158	0,61
7	introdução	274	58,36	7	linguístico	156	0,61
8	linguístico	273	58,14	8	educação	154	0,60
9	análise	273	58,14	9	texto	145	0,56
10	texto	271	57,72	10	brasileiro	145	0,56

Assim como no cluster 1, as palavras de maior frequência no cluster 2 também são bastante semelhantes aos itens do corpus, de 10 lemas levantados apenas 1 diferente entre o corpus e o C2, os demais itens apresentam semelhanças, mudando apenas as suas frequências de aparição, com relação ao valor absoluto. Sendo a palavra “introdução” que aparece no corpus com o valor absoluto de 274 e o termo “educação” que possui 154 como valor absoluto, essas palavras não são mencionadas nas duas tabelas. Já os demais termos aparecem em ambas, como é o caso das palavras “português” que aparece como o termo mais recorrente nas duas tabelas.

O termo “linguístico” possui uma frequência de 273 no corpus e 156 no C2, isso diz respeito a recorrência do termo linguístico nesses textos, em ambos os casos a palavra apresenta recorrência, mas ainda não se destaca comparada ao termo “literatura”. Pois as palavras que possuem uma maior frequência na comparação do C2 com o corpus são as palavras que estão

diretamente relacionadas a português, literatura, língua, ensino e estudos. Assim, não é possível identificar a frequência dos termos relacionados a linguística nesse caso.

Como pode-se ver abaixo, o termo Linguística não aparece nos bigramas das palavras de maior frequência no corpus e no cluster 2, esse termo não é citado como um dos dez bigramas de destaque. Desse modo, comprova o que foi analisado na tabela 5, que o termo “linguística” não é algo tão recorrente nos textos analisados. Os bigramas em destaques abordam o ensino da língua portuguesa, as teorias literárias mais recorrentes e a relação das práticas pedagógicas adotadas nas formações de professores.

Quanto ao C2, em comparação à amostra total, os bigramas mais recorrentes estão descritos abaixo:

Tabela 6: Dez bigramas (tokens) mais frequentes no Cluster 2, em comparação ao corpus total

CORPUS				C2			
Posição	Bigrama	Valores absolutos	Valores percentuais	Posição	Bigrama	Valores absolutos	Valores percentuais
1	língua portuguesa	259	0,55	1	língua portuguesa	112	0,44
2	literatura portuguesa	131	0,28	2	ensino fundamental	55	0,21
3	literatura brasileira	130	0,28	3	literatura portuguesa	54	0,21
4	ensino fundamental	75	0,17	4	prática pedagógica	47	0,18
5	prática pedagógica	58	0,13	5	literatura brasileira	37	0,14
6	estudos literários	49	0,09	6	estudos literários	24	0,09
7	língua latina	41	0,08	7	língua latina	21	0,08
8	português brasileiro	40	0,08	8	gramática latina	18	0,07
9	ensino médio	34	0,07	9	educação básica	18	0,07
10	educação básica	34	0,07	10	teoria literária	17	0,07

No que diz respeito aos bigramas com o item “linguística”, o de maior destaque no *corpus* é o bigrama linguística geral, que se refere ao livro de Ferdinand de Saussure que já foi falado acima, é uma das principais referências ao falar sobre linguística, desse modo as instituições sempre se utilizam dele como referência nas disciplinas que diz respeito as teorias linguísticas. No corpus ele aparece na primeira posição como sendo o bigrama de linguística mais citado, já no C2 ele aparece como a segunda posição, tendo os valores absolutos de 27 e 9.

Em seguida, as palavras que aparecem com a maior frequência juntas, são: linguística textual, linguística aplicada, linguística histórica. Esses são alguns dos títulos que recorrentemente são dados a componentes curriculares dos cursos, por isso a frequência nas suas aparições. Também aparecem termos como variação linguística, que chega a aparecer como tópicos das ementas, já que faz parte de um dos conteúdos mais abordados nos primeiros semestres das graduações e também aparecem com frequência em títulos de livros.

A extração de bigramas com o item linguística no C2, em contraponto ao corpus total, vê-se em detalhe abaixo:

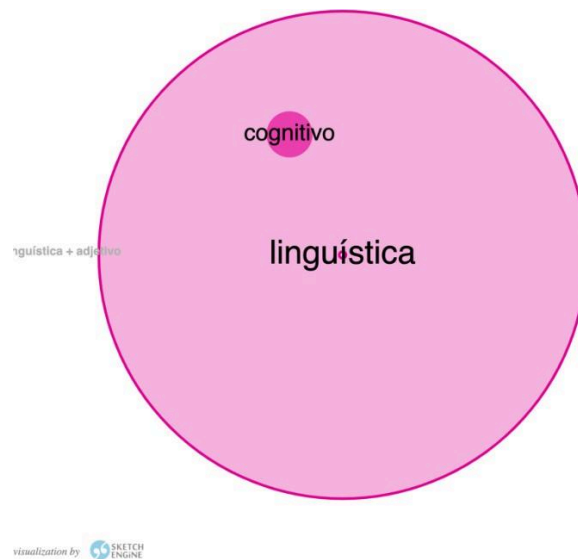
Tabela 7: Dez bigramas (tokens) com o item *linguística* mais frequentes no Cluster 2, em comparação ao corpus total

CORPUS				C2			
Posição	Bigrama	Valores absolutos	Valores percentuais	Posição	Bigrama	Valores absolutos	Valores percentuais
1	linguística geral	27	0,06	1	linguística aplicada	10	0,04
2	linguística textual	21	0,04	2	linguística geral	9	0,03
3	linguística aplicada	19	0,04	3	análise linguística	9	0,03
4	análise linguística	18	0,04	4	linguística textual	8	0,03
5	linguística descritiva	8	0,02	5	linguística histórica	6	0,02
6	linguística histórica	7	0,01	6	linguística românica	3	0,01
7	mudança linguística	7	0,01	7	política linguística	3	0,01
8	variação linguística	6	0,01	8	mudança linguística	3	0,01
9	linguística teórica	3	0,01	9	variação linguística	3	0,01
10	linguística cognitiva	3	0,01	10	linguística funcional	2	0,01

No Cluster 2 o bigrama de linguística de maior recorrência é “linguística aplicada” com 10 de valor absoluto e linguística funcional como o bigrama de menor recorrência, apenas com 2 no total do valor absoluto, esse bigrama não chega a aparecer no levantamento do corpus. Assim, pode-se notar que as teorias linguísticas funcionais são abordadas com uma frequência menor nos textos das universidades da classe 2 e quando falamos no corpus geral, esse bigrama não faz parte das dez posições mais frequentes.

A rede de colocados com o nome linguística como nó, vinculado a adjetivos, se mostra na figura abaixo:

Figura 3: Rede de colocados com o item linguística no C2



Ao falar na comparação entre a rede de colocados que mais se destaca entre o corpus e o Cluster 2, apenas o termo cognitivo aparece sendo ligado a linguística, dessa forma, nessa relação entre os textos predomina apenas a abordagem da linguística cognitiva.

4.3. Cluster 3

Por fim, apresentam-se os resultados quanto ao cluster 3, de que fazem parte dados textuais dos PPCs da UFPEL, UFM, FURG, UNIPAMPA, UFG e UNIFESP. A extração das 20 palavras (lemas) mais frequentes se detalha abaixo em comparação com as dez palavras mais frequentes no corpus total:

Tabela 8: Dez palavras (lemas) mais frequentes no Cluster 3, em comparação ao corpus total

CORPUS				C3			
Posição	Item	Valores absolutos	Valores percentuais	Posição	Item	Valores absolutos	Valores percentuais
1	português	775	165,06	1	português	324	1,87
2	literatura	720	153,34	2	literatura	307	1,77
3	língua	519	110,54	3	língua	202	1,17
4	ensino	402	85,62	4	ensino	163	0,94
5	estudo	359	76,46	5	brasileiro	157	0,91
6	brasileiro	334	71,14	6	introdução	137	0,79
7	introdução	274	58,36	7	estudo	122	0,70
8	linguístico	273	58,14	8	educação	117	0,67

9	análise	273	58,14	9	texto	115	0,66
10	texto	271	57,72	10	gramática	109	0,63

Dentre os Cluster criados esse é o único que o resultado das dez palavras mais frequentes do corpus e do C3 aparecem nas mesmas posições, mudando apenas os valores da frequência, uma vez que os valores do corpus são maiores do que os dos valores do C3. As palavras de maior frequência nesse cluster são os termos já falados nas seções anteriores, pois ele se assemelha bastante com o corpus geral, até no que diz respeito aos valores absolutos, apresentando o valor absoluto de 775 como frequência da palavra “português”, que assume a primeira posição das duas tabelas. Palavra essa que no bigrama abaixo aparece relacionada ao termo língua, assim explicando a causa de ser o bigrama mais frequente, já que aparece em quase todas as ementas, títulos das disciplinas e das obras que fazem parte da biografia, por trata-se de cursos dessa língua.

A palavra “linguístico” aparece na posição de número oito, com um volume absoluto da frequência de 273 tanto no corpus como no C3. Portanto, esse termo não possui uma predominância nos textos referidos, possuindo uma frequência de aparições menor do que alguns termos como, “literatura”, “língua”, “ensino”, “estudo” e entre outros. Ao comparar com os termos destacados quando é levantado o bigrama dos termos de maior frequência, o bigrama de linguística não aparece no corpus, aparecendo apenas no C3, na posição nove com os valores absolutos de 18 e o termo aparece como “linguística geral”, fazendo referência ao livro já destacado nos tópicos anteriores.

Do cluster 3 foram extraídos os seguintes bigramas mais recorrentes:

Tabela 9: Dez bigramas (tokens) mais frequentes no Cluster 3, em comparação ao corpus total

CORPUS				C3			
Posição	Bigrama	Valores absolutos	Valores percentuais	Posição	Bigrama	Valores absolutos	Valores percentuais
1	língua português	259	0,55	1	língua português	103	0,59
2	literatura português	132	0,28	2	literatura brasileiro	72	0,42
3	literatura brasileiro	130	0,28	3	literatura português	54	0,31
4	ensino fundamental	79	0,17	4	português brasileiro	25	0,14
5	prática pedagógico	60	0,13	5	ensino fundamental	21	0,12
6	português brasileiro	40	0,09	6	resolução cne	20	0,12

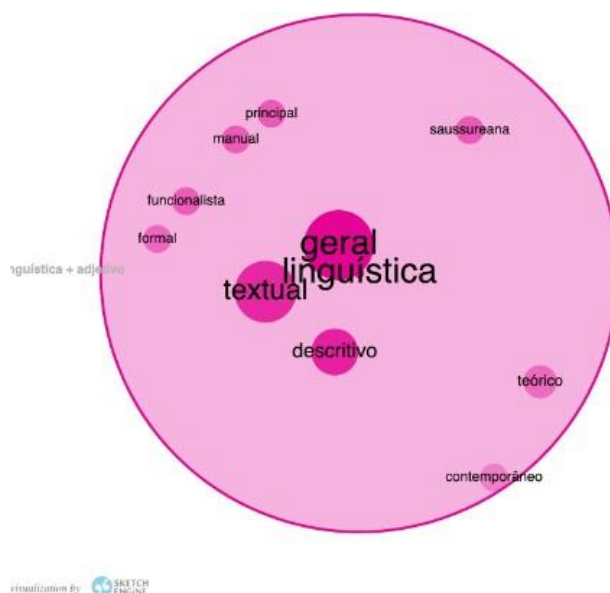
7	língua latino	39	0,08	7	ensino médio	18	0,10
8	estudo literário	36	0,08	8	produção textual	18	0,10
9	educação básico	34	0,07	9	linguística geral	18	0,10
10	ensino médio	34	0,07	10	língua latino	16	0,09

Os bigramas de maior frequência correspondem aos lemas que eram destaques na tabela anterior, destacando as primeiras posições dos bigramas do corpus e do C3, que são semelhantes, mudando apenas a posição, os destaques são: “língua português”, “literatura português” e “literatura brasileiro”. Logo, destaca a literatura como sendo a abordagem de maior destaque nos textos estudados, referindo-se a nomes de disciplinas que utilizam dessas abordagens, títulos de livros e tópicos das ementas dos componentes curriculares.

Tabela 10: Dez bigramas (tokens) com o item *linguística* mais frequentes no Cluster 3, em comparação ao corpus total

CORPUS				C3			
Posição	Bigrama	Valores absolutos	Valores percentuais	Posição	Bigrama	Valores absolutos	Valores percentuais
1	linguística geral	27	0,06	1	linguística geral	18	0,10
2	linguística textual	21	0,04	2	linguística textual	13	0,07
3	linguística aplicada	19	0,04	3	linguística aplicada	9	0,05
4	análise linguística	18	0,04	4	análise linguística	7	0,04
5	linguística descritiva	8	0,02	5	linguística descritiva	6	0,03
6	linguística histórica	7	0,01	6	mudança linguística	4	0,02
7	mudança linguística	7	0,01	7	variação linguística	3	0,02
8	variação linguística	6	0,01	8	linguística teórica	2	0,01
9	linguística teórica	3	0,01				
10	linguística cognitiva	3	0,01				

Figura 4: Rede de colocados com o item linguística no C3



5. CONSIDERAÇÕES FINAIS

Neste trabalho, os Projetos Pedagógicos de Cursos foram utilizados como corpus de análises, por tratar-se de documentos oficiais das instituições que têm o poder de mostrar como funciona a formação discente nos cursos de graduações, uma vez que, informa sobre quais teorias são repassados para os discentes no decorrer da sua formação e as referências bibliográficas utilizadas. Tornando-se o melhor objeto de estudo para conhecer como o termo *linguística* costuma ser referido nos PPCs dos cursos de Letras-Português das universidades federais.

Partindo das análises, conclui-se que a palavra “linguística” e os termos relacionados, apareceram com menos frequência nos textos dos agrupamentos do que no corpus total. Já os termos que se destacam no corpus total e nos clusters analisados, foram “português” e “literatura”. Seguido de lemas ligados aos estudos literários, destacando os títulos de algumas das disciplinas que aparecem com frequências nos diferentes cursos, como é o caso da “Literatura Brasileira”, “Literatura Portuguesa” e “Estudo literário”.

As maiores frequências de “linguística” se deram na extração dos bigramas, os quais destacaram o termo em conjunto com a palavra geral, formando assim “linguística geral”, fazendo referência a obra *Curso de Linguística Geral* de Ferdinand Saussure, que pode então ser considerado como o principal material de referência ao tratar das teorias linguísticas.

Com isso, foi possível identificar quais as universidades que mais se assemelham através dos seus PPCs, que foram identificadas por meio dos clusters que se formaram nos dendrogramas. E Através das análises dos termos de maior recorrência nesses clusters, é notório que os lemas que mais coocorrem com a linguística, estão vinculadas as vertentes

educacionais da linguística, como é o caso de linguística funcional, cognitiva, textual, descritiva e entre outras.

Com base nos dados expostos, pode-se notar que durante a formação do curso de Letras-Português das universidades federais o eixo de formação vinculado a linguística não prevalece, diferentemente da formação literária, já que esse léxico se sobressai. Portanto, o que se tentou fazer nesse trabalho, foi identificar por meio de um estudo de colocados se as teorias linguísticas adotadas nas formações docentes dos cursos de Letras-Português das universidades federais se assemelham.

6. REFERÊNCIAS

BAKER, P.; HARDIE, A.; MCENERY, T. **A Glossary of Corpus Linguistics**. Edinburgh: Edinburgh University Press, 2006.

BORTONI-RICARDO, Stella Maris & FREITAS, Vera Aparecida de Lucas. **Sociolinguística Educacional**. In: ABRALIN: 40 ANOS EM CENA. João Pessoa: Editora Universitária, 2009.

BRASIL. Ministério da Educação. Diretrizes Curriculares Nacionais dos cursos de Filosofia, História, Geografia, Serviço Social, Comunicação Social, Ciências Sociais, Letras, Biblioteconomia, Arquivologia e Museologia. Parecer CNE/CES 492/2001, homologação publicada no DOU 9/7/2001, Seção 1e, p. 50. Resolução CNE/CES 18/2002, DOU 09/04/2002, Seção 1, p. 34.

BRASIL. Lei nº 10.861, de 14 de abril de 2004. Institui o Sistema Nacional de Avaliação da Educação Superior – SINAES. Brasília, 2004.

BRASIL: Ministério da Educação. Ministério da Educação. Disponível em: <http://portal.mec.gov.br/component/tags/tag/32911>. Acesso em: 12 nov. 2022.

BRASIL. **Resolução CNE/CP n.2, de 20 de dezembro de 2019**. Define as Diretrizes Curriculares Nacionais para a Formação Inicial de Professores para a Educação Básica e institui a Base Nacional Comum para a Formação Inicial de Professores da Educação Básica (BNC-Formação), 2019b.

CANTOS GÓMEZ, Pascual. Do we need statistics when we have linguistics?. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 18, p. 233-271, 2002.

DEMŠAR, Janez et al. Orange: data mining toolbox in Python. **The Journal of machine Learning research**, v. 14, n. 1, p. 2349-2353, 2013.

DONI, Marcelo Viana. **Análise de cluster: métodos hierárquicos e de particionamento**. Universidade Presbiteriana Mackenzie, 2004.

DUTT-ROSS, S.; CRUZ, B. P. A. **Análise Quantitativa de Textos: Apresentação e Operacionalização da Técnica via Linguagem R no Twitter**. Administração: Ensino e Pesquisa, v. 22, n. 1, p. 186-220, 2021.

FIORIN, J. L. (Org.). **Introdução à linguística: objetos teóricos**. 6. ed. São Paulo: Contexto, 2010.

GIL, Antonio C. **Como Elaborar Projetos de Pesquisa**, 7ª Edição. Atlas, 2022.

KILGARRIFF, A. *et al.* The Sketch Engine: ten years on. **Lexicography**, v. 1, n. 1, p. 7–36, 1 jul. 2014.

MUSSALIM, Fernanda; Bentes, Anna Christina. **Introdução á linguística: domínios e fronteiras**. São Paulo: Editora Cortez, 2009.

OLIVEIRA, Lucy. Análise de texto automatizada e análise de conteúdo: abordagens combinadas e apontamentos sobre a produção latino-americana. In: ASSOCIAÇÃO LATINO-AMERICANA DE CIÊNCIAS POLÍTICAS (ALACIP), 2019. Monterrey, México. Anais... Disponível em: <https://alacip.org/cong19/241-oliveira-19.pdf>. Acesso em: 10 nov. 2022.

OLIVEIRA, Sheila Elias de. **O PAPEL DA LINGÜÍSTICA NA FORMAÇÃO DO PROFESSOR DE LÍNGUA**. Conexão Letras, Paraná, v. 2, n. 2, p. 73-85, 2006.

ORLANDI, Eni. **O que é lingüística**. São Paulo: Editora Brasiliense, 1992.

ROBERTS, Carl W. A conceptual framework for quantitative text analysis. **Quality and Quantity**, v. 34, n. 3, p. 259-274, 2000.

SEIXAS, Pablo Sousa; COELHO-LIMA, Fellipe; SILVA, Suzany Gadelha; YAMAMOTO, Oswaldo Hajime. **Projeto Pedagógico de Curso e formação do psicólogo: uma proposta de análise**. Revista Semestral da Associação Brasileira de Psicologia Escolar e Educacional, São Paulo, v. 17, n. 1, p. 113-122, jun. 2013.

SVARTVIK, J. **Corpora are becoming mainstream**. In: THOMAS, J. and SHORT, M. (orgs). Using corpora for language research. London and New York: Longman, 1996. p 3-13.