Entretextos

ENTRETEXTOS. Londrina, v. 24, n.1, 2024.
ISSN 1519-5392 UEL
DOI: 10.5433/1519-5392.2024v24n1p106-130

Linguística de Corpus e *Sketch Engine*: a seleção lexical do Projeto de Assentamento São Francisco

Corpus Linguistics and Sketch Engine: the lexical selection of the PA São Francisco Settlement Project

Lingüística de Corpus y Sketch Engine: la selección léxica del Proyecto de Asentamiento São Francisco

Andreza Marcião dos Santos[1]
0000-0002-9777-2829
Maria Cândida Trindade Costa de Seabra[2]
0000-0003-4827-0635

RESUMO: Este artigo visa a apresentar uma reflexão sobre a importância da Linguística de Corpus (LC) como aporte metodológico para a seleção lexical do *corpus* oral do Projeto de Assentamento São Francisco, com o auxílio do programa computacional Sketch Engine. Nesse viés, destaca-se a LC como uma metodologia que possibilita a análise de dados da língua de forma probabilística e que permite analisar padrões ou tendências de um fenômeno linguístico. Além disso, está ligada à criação de *corpora* eletrônicos e a utilização de softwares de análise para a leitura de corpus oral ou escrito. O *Sketch Engine*, por exemplo, permite a comparação entre um *corpus* de estudo e um corpus de referência, de modo a destacar as palavras-chave por meio da análise de frequências em diferentes contextos. Ressalta-se, portanto, a proposta metodológica de constituição de um corpus oral e o processamento de dados para futuras análises de cunho lexical, que visam explorar dados linguísticos de uma língua, especialmente em contextos específicos de pesquisa linguística.
PALAVRAS-CHAVE: linguística de corpus; *sketch engine*; léxico.

ABSTRACT: This article aims to present a reflection on the importance of Corpus Linguistics (LC) as a methodological contribution for the lexical selection of the oral corpus of PA São Francisco, with the help of the computer program Sketch Engine. In this sense, LC is highlighted as a methodology that enables the analysis of language data in a probabilistic way and that allows analyzing patterns or trends in linguistic phenomen. Furthermore, it is linked to the creation of electronic *corpora* and the use of analysis software for reading both oral and written *corpora*. For example, Sketch Engine allows the comparison between a corpus of study and a reference corpus in order to highlight keywords through frequency

[1] Doutora em Estudos Linguísticos pela Universidade Federal de Minas Gerais - UFMG. Residência Pós-Doutoral em Estudos Linguísticos pela UFMG. E-mail: andrezamarciao@hotmail.com
[2] Doutora em Estudos Linguísticos pela Universidade Federal de Minas Gerais - UFMG. Professora Titular da Faculdade de Letras – FALE/UFMG. E-mail: candidaseabra@gmail.com

analysis in different contexts. Thus, the methodological proposal of constructing an oral corpus and processing this data for future lexical analyses aimed at exploring linguistic data in a language, especially in specific linguistic research contexts.
KEYWORDS: corpus linguistics; sketch engine; lexicon.

RESUMEN: Este artículo tiene como objetivo presentar una reflexión sobre la importancia de la Lingüística de Corpus (LC) como contribución metodológica para la selección léxica del corpus oral de la PA São Francisco, con la ayuda del programa informático Sketch Engine. En este sentido, LC destaca como una metodología que posibilita el análisis de datos lingüísticos de forma probabilística y que permite analizar patrones o tendencias en un fenómeno lingüístico. Además, está vinculado a la creación de corpus electrónicos y al uso de software de análisis para leer corpus orales o escritos. Sketch Engine, por ejemplo, permite comparar entre un corpus de estudio y un corpus de referencia, para resaltar palabras clave mediante análisis de frecuencia en diferentes contextos. De esta manera, se destaca la propuesta metodológica de constituir un corpus oral y procesar estos datos para futuros análisis léxicos que tengan como objetivo explorar datos lingüísticos de una lengua, especialmente en contextos específicos de investigación lingüística.
PALABRAS CLAVE: lingüística de corpus; sketch engine; léxico.

Introdução

A Linguística de Corpus (doravante LC) é compreendida, neste estudo, como uma abordagem teórico-metodológica de investigação linguística sobre o uso da língua que possibilita fazer descrições e generalizações de uma língua de uma forma empírica, por meio de programas computacionais. Segundo Sardinha (2004), a LC pode auxiliar nos estudos da lexicogramática, léxico de forma probabilística, ou seja, verificando as regularidades e padronizações que podem indicar características de uma determinada língua.

Essa metodologia também orienta os passos para a constituição de um *corpus/corpora* oral ou escrito, a partir de um conjunto de textos escritos ou transcrições orais, em formato eletrônico. Segundo McEnery e Hardie (2012), a LC possui o potencial de fornecer informações sobre os fenômenos linguísticos presentes na linguagem escrita e oral, mas, também, permite visualizar com que frequência eles ocorrem, tornando-se uma ferramenta importante para a análise do que é possível ou do que é provável em uma língua.

O *corpus* de análise geralmente apresenta duas maneiras de mostrar os resultados, sendo: a concordância ou a frequência. De acordo com Lindquist e Levin (2009), a concordância é uma lista de todos os contextos em que uma

palavra ocorre, tornando-se um material para pesquisadores extraírem informações lexicais ou de outros níveis da língua. A frequência, também, é um fator importante para verificar a regularização de fenômenos linguísticos que pode estar presente no *corpus*, sendo facilmente obtida por meio de programas computacionais. Esses processos permitem a descrição e a identificação das diferenças de gêneros, variedades geográficas, língua falada ou escrita, bem como textos de diferentes períodos de tempo.

É importante ressaltar que a LC não se restringe à observação de como os falantes utilizam a língua no ato comunicativo, mas também oferece possibilidades para compreender e descrever os usos da língua. Nesse contexto, a constituição de um *corpus/corpora*, associada ao processamento computacional, surge como um processo essencial para a realização dessa tarefa. Conforme destaca Fillmore (1992), embora não exista um *corpus* que contenha todas as informações de uma língua, cada *corpus* tem o potencial de revelar aspectos linguísticos relevantes sobre uma dada língua. Dessa forma, consideramos o Projeto de Assentamento São Francisco (doravante PA São Francisco) para a constituição de um *corpus* de estudo, de modo a compreender como os dados são tratados e processados pelos programas computacionais, com o aporte da LC.

Sendo assim, este texto apresentará como o programa *Sketch Engine* e suas ferramentas desempenham um papel importante no processamento e organização desses dados, principalmente quando se trata de um *corpus* oral. Contudo, ele não determina por si só a análise da realidade linguística dos sujeitos. Em outras palavras, a interpretação e a atribuição de significado aos dados processados dependem do pesquisador, pois é ele que envolverá as perspectivas teóricas e abordagens específicas para analisar os dados processados pelo programa.

Os fundamentos teórico-metodológicos da Linguística de Corpus

A Linguística de corpus (LC) se ocupa da coleta e exploração de *corpora* ou de um conjunto de dados linguísticos que foram coletados de forma criteriosa com o propósito de servir para análise de uma língua ou variedade linguística. Esse

processo de pesquisa sobre a língua também envolve o uso de *softwares* computacionais (Sardinha, 2000).

Segundo Sardinha (2004), muitos pesquisadores, em boa parte do século XX, se dedicaram à descrição da língua por meio de *corpora*. Entretanto, neste período, os *corpora* não eram publicados no formato eletrônico, mas, sim, coletados e analisados manualmente. Destaca-se como *corpus* não eletrônico, o *Survey of English Usage* (SEU)[2], compilado por Randolf Quirk e sua equipe, a partir de 1959.

O SEU serviu de base para a constituição do primeiro *corpus* linguístico eletrônico, denominado de *corpus Brown.* Esse *corpus* foi lançado em 1964 com a quantidade de 1 milhão de palavras, sendo que a sua posição de destaque não se deu somente por essa característica, mas também pela época em que gastar tempo e recursos financeiros para a coleta de registros era visto com incredulidade e hostilidade (Sardinha, 2004). De fato, realizar a coleta de dados linguísticos e, posteriormente, acoplá-los em um banco de dados requer tempo, recursos financeiros e humanos, além de conhecimentos acerca dos programas que permitem a leitura e o processamento dos dados coletados.

Por essa razão, não há como falar de *corpus* eletrônico sem considerar a invenção dos computadores e o desenvolvimento tecnológico que possibilitam o acesso, compartilhamento, processamento e análise de *corpora*. Consequentemente, permite-se a criação e a manutenção de *corpora* cada vez maiores em número de palavras, visto que a LC também se vincula e está condicionada à tecnologia.

Nesse sentido, compreende-se que um *corpus* se constitui, segundo Sardinha (2004, p. 18), "como um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou ambos), de maneira que sejam representativos da totalidade do uso linguístico ou que propiciem resultados vários e úteis para a descrição e análise". Um *corpus* deve ser planejado e concretizado seguindo alguns critérios, sendo eles: i) a origem, os dados devem ser autênticos; ii) o propósito, o corpus deve ter a finalidade de ser um objeto de estudo linguístico[2]; iii) a composição, o

[2] Entende-se que o *corpus* não deve ser apenas um repositório aleatório de dados linguísticos, mas deve ser criado com a intenção de ser examinado e analisado sob o viés linguístico. É uma forma de investigar e compreender aspectos específicos da língua, como, por exemplo, padrões sintáticos,

conteúdo do *corpus* deve ser criteriosamente escolhido[3]; iv) a formatação, os dados do *corpus* devem ser legíveis em computador (formato. txt); v) a representatividade, o *corpus* deve ser representativo de uma língua ou variedade; vi) a extensão, o *corpus* deve ser vasto para ser representativo.

Cabe destacar que a representatividade, neste texto, não é compreendida em termos de um *corpus* que possui uma grande quantidade de palavras de uma determinada língua ou variedade, mas, sim, como um *corpus* em que conste o maior número possível de acepções de cada forma. Por exemplo, a forma *como* pode significar uma preposição ou a primeira pessoa do singular do verbo comer no presente do indicativo. Deve-se, portanto, verificar quais formas estão presentes no *corpus* e como isso contribui para o estudo e conhecimento da língua.

Nesse caso, é possível estudar a língua de forma probabilística e verificar as regularidades e padronizações que podem indicar as características de uma determinada variedade. Além disso, é necessário que o pesquisador tenha conhecimento acerca das ferramentas eletrônicas e de como utilizá-las para o processo de análise dos fenômenos apresentados no *corpus*. Contudo, vale ressaltar que dominar as ferramentas eletrônicas, não é suficiente. É fundamental que o pesquisador também apresente um conhecimento linguístico significativo, para poder delimitar os caminhos teórico-metodológicos de leitura e análise, visto que ter acesso aos dados é pouco eficaz se não houver a habilidade de interpretá-los, conforme o objetivo da pesquisa.

Segundo Davies (2015), alguns dos fenômenos que podem ser estudados em um *corpus* envolvem o nível lexical (frequência e distribuição de palavras específicas ou frases, lista de todas as palavras comuns na língua ou gênero), morfológico (processos envolvendo a formação de palavras), padrões fraseológicos (preferências de colocação para palavras específicas), semântico (por exemplo, semelhança entre palavras, traços semânticos opostos e polissemia). Essas possibilidades podem produzir *insights* para a compreensão de como a língua é

variações linguísticas e fenômenos lexicais.

[3] Este critério envolve a naturalidade e autenticidade do *corpus*. Por isso, a seleção dos textos, ou dados do *corpus,* deve ser feita de acordo com critérios específicos, como a inclusão de fatores de gênero textual, período de texto e contexto social, o que garantirá que o *corpus* atenda aos objetivos específicos de uma dada pesquisa.

utilizada por seus usuários.

Assim, se o pesquisador optar por coletar dados e construir o seu próprio *corpus*, ele deve considerar alguns pontos. No caso do *corpus* escrito, Rayson (2015) afirma que ao compilá-lo é necessário considerar o material impresso, digitalizar materiais antigos ou escritos à mão, converter para o formato. txt; utilizar, quando necessário, ferramentas que coletam dados como o *BootCat, WebBooCat e Sketch Engine*. Para o *corpus* oral, deve-se observar: i) as gravações devem ter alta qualidade; ii) utilização de *software* de transcrição e edição, como o *VoiceWalker*; iii) utilização de *softwares* em análise e síntese da fala, como o PRAAT; iv) alinhamento de áudio e vídeo.

Pela perspectiva metodológica da LC, é possível considerar tanto a modalidade escrita quanto a modalidade oral da língua, para constituir diferentes *corpora* e descobrir novas formas de analisar e estudar os fenômenos da língua. Estudos recentes de Sardinha *et al.* (2022), apontam a modalidade visual na vertente da Análise Multidimensional Imagética (AMDI) para analisar imagens presentes em tweets, com base na LC. Isso contribui para a análise dos eventos comunicativos realizados por texto-imagens, ou seja, um novo caminho que vai além dos textos orais e escritos. Sendo assim, a elaboração de um *corpus* envolve a escolha inicial de uma modalidade da língua, e, posteriormente, o prosseguimento com as investigações acerca da língua (Biber, 1993).

Sardinha (2000, p. 341) explica que o *corpus* apresenta duas finalidades, sendo elas: um *corpus* de estudo e um *corpus* de referência. O primeiro se refere ao *corpus* que se pretende descrever, ou seja, a coleção de textos que o pesquisador está investigando de forma específica. Os critérios e as fontes de coleta de dados, por exemplo, são selecionados para atender aos objetivos da pesquisa, sendo que, por meio desse *corpus* de estudo, identificam-se os padrões linguísticos, não somente tendências de uma língua, ou uma variedade linguística. Enquanto o segundo é usado para fins de contraste com o *corpus* de estudo. Isso ocorre porque o *corpus* de referência possui uma coleção maior de textos e contempla diferentes gêneros ou registros, estilos e fontes[4] da diversidade de uma língua.

[4] Significa que o *corpus* de referência é composto por uma grande quantidade de textos provenientes

Frequentemente, é utilizado para estabelecer comparações entre o contexto mais geral da língua e o contexto mais específico da língua (*corpus* de estudo) que está sendo estudado.

Entende-se que ambos são ferramentas disponibilizadas pela LC para entender como a língua é usada pelos falantes e como os padrões linguísticos podem variar de acordo com os contextos de produção. Assim, retoma-se ao *corpus* oral, que permite a reflexão sobre como o falante utiliza a língua em seu cotidiano, considerando as situações autênticas de comunicação. Destaca-se que textos escritos também apresentam esta característica, pois é produzido por falantes que desejam transmitir informações em contextos específicos.

Segundo Biber (1998), há uma correlação entre características linguísticas e os contextos de uso da língua devido aos conjuntos de traços linguísticos que variam sistematicamente a partir desses dois fatores, concluindo que a variação não é aleatória, mas padronizada conforme a evidência de recorrência dos fenômenos a serem analisados. Em outras palavras:

> A linguagem forma padrões que apresentam regularidade (se mostram estáveis em momentos distintos, isto é, tem frequência comparável em corpora distintos) e variação sistemática (correlacionam-se com variedades textuais, genéricas, dialetais, etc). Exemplos notáveis da descrição da linguagem por meio da indução de padrões recorrentes são a gramática de verbos (Francis, G. e Hunston, 1996) e de substantivos e adjetivos (Francis, G. e Hunston, 1998) lançadas pelo projeto COBUILD, nas quais se descreve exaustivamente todos os padrões lexicais existentes na língua inglesa (Sardinha, 2000, p. 351).

Nesse sentido, as possibilidades de uso da língua apresentadas pelo falante incorporam as probabilidades de ocorrência de traços linguísticos, que podem auxiliar na compreensão das estruturas linguísticas a partir da observação dos dados coletados. Assim, é importante considerar um *corpus* oral como uma fonte de informação, de registro natural na língua e de verificação das frequências de ocorrência, pois podem atestar realidades não somente linguísticas, mas também realidades demográficas, culturais e sociais.

Dessa forma, a próxima seção explica como o *corpus* de estudo foi utilizado

---

de diversas fontes, isto é, se refere às origens dos textos, que podem incluir textos literários e conversações cotidianas, entre outros contextos de produção de linguagem.

nesta pesquisa e como se deu o processo de extração das palavras-chave do léxico dos falantes do PA São Francisco, na tentativa de empreender análises para o conhecimento de uma realidade que envolve diferentes sujeitos (oriundos de diversas regiões brasileiras) e de ampliação do conhecimento da língua, a partir da descrição de uma variedade ainda não descrita.

A constituição do corpus oral do PA São Francisco

O *corpus* oral do PA São Francisco foi elaborado e coletado pelas autoras nos anos de 2020 e 2021[5], com a finalidade de realizar a descrição do léxico dessa comunidade localizada no sul do estado do Amazonas e de proporcionar uma fonte de referência para outras pesquisas no contexto amazônico. A construção do *corpus* se deu por meio de entrevistas com os assentados em suas residências, totalizando 13 sujeitos-agentes. O Quadro 1 mostra a organização dos metadados[6] em gênero, faixa etária e tempo de gravação individual e total.

[5] A pesquisa foi submetida e aprovada pelo Comitê de Ética da Plataforma Brasil, sob o CAAE: 43158021.9.0000.5149.
[6] Os metadados são variáveis independentes (fatores extralinguísticos) que ajudam na explicação da variação.

Quadro 1 – Os sujeitos-agentes do corpus oral do PA São Francisco

Faixa etária	**18 a 35 anos**[7]				
Número da entrevista	01	02	05	11	
Gênero	M	M	M	M	
Idade	35	27	18	24	
Duração da entrevista	46:56	20:36	26:38	29:17	
Duração total das entrevistas	2 horas, 03 minutos e 45 segundos				
Faixa etária	**36 a 55 anos**				
Número da entrevista	03	09	12	13	
Gênero	F	M	M	F	
Idade	41	46	50	42	
Duração da entrevista	60:00	25:39	35:19	46:26	
Duração total das entrevistas	2 horas, 47 minutos e 04 segundos				
Faixa etária	**56 em diante**				
Número da entrevista	04	06	07	08	10
Gênero	F	M	F	M	M
Idade	61	67	62	64	58
Duração da entrevista	44:09	56:39	35:54	56:41	39:40
Duração total das entrevistas	3 horas, 53 minutos e 03 segundos				

Fonte: Elaborado pelas autoras.

A opção por essas faixas etárias se deu pelo registro de fala de diferentes gerações e de como cada informante utiliza o léxico em diferentes contextualizações temáticas. Nesse viés, percebemos que as diferentes faixas etárias proporcionaram tempos de gravação distintas, pois, os mais jovens apresentaram uma fala mais sucinta, sendo finalizadas em menos de 30 minutos, enquanto os mais velhos eram mais detalhistas, atingindo o tempo de até 60 minutos de gravação.

Outro aspecto importante a ser ressaltado é a característica individual de cada entrevistado, sendo que, para conhecer um pouco mais sobre os sujeitos-agentes, foi organizado o Quadro 2, com o informativo sobre o perfil de cada um.

[7] O eixo diagenérico foi considerado durante a coleta de dados e a ausência de sujeitos-agentes femininos nesta faixa etária de 18 a 35 anos também foi observada. No entanto, um dos fatores que contribuiu para essa lacuna foi o difícil acesso às casas dos entrevistados, visto que é uma área de assentamento de reforma agrária e as estradas não são asfaltadas.

Quadro 2 – Perfil dos sujeitos-agentes

CÓDIGO DA ENTREVISTA	ESTADO DE ORIGEM	OCUPAÇÃO	ESCOLARIDADE	TEMPO DE RESIDÊNCIA NO PA
01HM35	RO	Serviços gerais/agricultor	Ensino fundamental I	10 anos
02JM27	RO	Agricultor	Ensino Médio	5[8] meses
03LF41	PI	Agricultora	Ensino fundamental I	5 anos
04MF61	SP	Doceira	Analfabeta	12 anos
05AM18	RO	Estudante	Ensino médio	9 anos
06SM67	PE	Aposentado/ agricultor	Ensino fundamental I	16 anos
07JF62	MT	Dona de casa	Ensino fundamental I	19 anos
08JM64	MS	Lavrador	Ensino fundamental I	13 anos
09GM46	MT	Lavrador	Ensino fundamental I	11 anos
10PM58	ES	Agricultor	Ensino fundamental I	24 anos
11PM24	MA	Faz tudo	Ensino médio incompleto	3 anos
12JM50	SC	Agricultor	Ensino fundamental I	20 anos
13SF42	RO	Agricultora	Ensino médio tecnológico	3 anos

Fonte: Elaborado pelas autoras.

O *código da entrevista* serviu para a identificação e a localização dos excertos extraídos das entrevistas dos assentados e para preservar a identidade dos participantes da pesquisa. A estruturação que gerou cada código de entrevista envolveu a seguinte padronização: 1) número da entrevista, considerando a ordem em que cada assentado foi entrevistado; 2) inicial dos nomes dos assentados; 3) gênero; e 4) faixa etária.

A identificação do *estado de origem* colabora para a compreensão da heterogeneidade linguística e também sociocultural dos entrevistados em um espaço territorial e geográfico situado no estado do Amazonas, do qual muito se fala, mas que ainda é pouco conhecido. A *ocupação* indica o trabalho realizado por cada entrevistado, sendo a maioria envolvido com a agricultura. O *nível de escolaridade* mostra o grau de instrução formal de cada participante da pesquisa, dando destaque para o Ensino Fundamental I. Por fim, o *tempo de residência no assentamento* contribui para a interpretação dos dados para o processo de criação, infraestrutura, vivências e dificuldades encontradas no PA São Francisco.

[8] Ressaltamos que, por um lado, o período de 5 meses pode ser considerado relativamente curto para se obter uma representação completa e estável do PA São Francisco. Por outro lado, consideramos este tempo porque foi possível captar uma gama significativa de expressões linguísticas, especialmente no que se refere à variação linguística.

Após as informações do perfil dos sujeitos-agentes, o próximo ponto foi a elaboração do roteiro de entrevista com as questões direcionadas e não direcionadas. O roteiro foi constituído com base no Questionário Semântico-Lexical (QSL), desenvolvido pelo Projeto Atlas Linguístico do Brasil (ALiB), e os questionários de Corrêa (1980), Cruz (2004), Azevedo (2013) e Batista (2019). Sendo assim, a sua estrutura foi formada por 85 questões, distribuídas em 8 campos semânticos.

Os campos semânticos foram selecionados a partir de uma análise dos campos semânticos apresentados no QSL do ALiB e no QSL de Cruz (2004). Ambos apresentaram questões sobre o léxico, conforme o Quadro 3.

Quadro 3 – Campos semânticos

ALiB	**Cruz (2004)**
a) Acidentes geográficos b) Fenômenos atmosféricos c) Astros e tempo d) Atividades agropastoris e) Fauna f) Corpo humano g) Ciclos da vida h) Convívio e comportamento social i) Religião e crenças j) Jogos e diversões infantis k) Habitação l) Alimentação e cozinha m) Vestuário e acessórios n) Vida urbana	**I) Meio físico** a) a terra e rios b) fenômenos atmosféricos **II) Meio biótico** a) fauna b) flora **III) Meio antrópico** a) o homem b) atividades de produção i) agricultura (roça, cultivo da juta, cultivo da mandioca) ii) caça e pesca iii) meios de transporte fluvial

Fonte: Elaborado pelas autoras.

Dentre os campos semânticos encontrados no ALiB, foram selecionados: fauna, convívio e comportamento social, religião e crenças, jogos e diversões infantis, com o objetivo de constituir campos semânticos mais gerais que possibilitam o conhecimento sobre o sujeito/informante, isto é, utilizados em outras pesquisas semântico-lexicais[9], independente dos estados brasileiros. Com relação aos campos semânticos de Cruz (2004), eles foram a base para a criação específica de outros

[9] Exemplos de pesquisa são: “Tabus Linguísticos nas capitais do Brasil: um estudo baseado em dados Geossociolinguísticos”, de Benke (2012); “A variação lexical no campo semântico vestuários e acessórios: um estudo a partir dos dados do Projeto ALiB, de Paim (2019) e “Os verbos botar e colocar no estado do Maranhão em dados do ALib: uma pesquisa variacionista, de Lavor, Araújo e Pereira (2020).

campos semânticos, a partir do conhecimento da pesquisadora sobre o PA São Francisco, como, por exemplo, os campos de assentamento, saúde, trabalho e meios de transporte, que abarcam o meio físico, meio biótico e meio antrópico, ao se pensar no contexto amazônico.

Assim, os campos semânticos selecionados e criados foram: 1) assentamento; 2) saúde; 3) convívio e comportamento social; 4) trabalho; 5) meios de transporte; 6) fauna; 7) religião e crenças; 8) jogos e diversões. Nas entrevistas os sujeitos-agentes responderam, por um lado, às questões direcionadas sobre os campos semânticos, por outro, questões não direcionadas para que, no momento de comunicação, houvesse uma interação mais natural e o informante ficasse mais à vontade para falar ou narrar fatos e experiências pessoais, de modo a não se monitorarem.

Posteriormente à realização das entrevistas, foram realizadas as transcrições para a formação do *corpus* de estudo. As normas de transcrição seguiram o modelo adotado no Projeto Filologia – "Pelas Trilhas de Minas: As Bandeiras e a Língua das Gerais[10]" (2003-2006). Com as entrevistas transcritas, constituiu-se um conjunto de dados em formato eletrônico, o que permitiu a leitura por computador. Ressalta-se que o processo de transcrição também envolveu o programa Praat[11] para a análise de voz, indicando a composição das unidades lexicais.

Após essa etapa, as transcrições foram salvas em formato.txt para a inserção no programa *Sketch Engine,* permitindo a leitura, processamento e análise dos dados. Conforme a caracterização proposta por Sardinha (2004), o *corpus* oral do PA São Francisco possui o número total de 53.941 *tokens*[12] e pode ser classificado como um *corpus* pequeno. O Quadro 4 apresenta, em dados numéricos, a extensão do *corpus* de estudo, considerando somente os sujeitos-agentes e outro constituído pela pesquisadora e os sujeitos-agentes.

[10] FAPEMIG – SHA844/02, coordenado pela Prof.ª Dr.ª Maria Antonieta Amarante de Mendonça Cohen.

[11] O Praat permite analisar características acústicas do som, como o tom da voz, intensidade (volume), duração e formantes. Essas análises são úteis para estudar aspectos da produção da fala, como entonação, ênfase e padrões de vogais e consoantes.

[12] Um *token* é caracterizado pelo número de itens identificados no corpus. Em outras palavras, é o número de ocorrências de uma palavra presente no *corpus.*

Quadro 4 – Extensão do *corpus* de estudo

***Corpus* de estudo**	**Total de palavras (*tokens)***
Sujeitos-agentes	53.941
Pesquisadora e sujeitos-agentes	78.591

Fonte: Elaborado pelas autoras.

Além da extensão e das informações sobre os sujeitos-agentes, o *corpus* possui as seguintes características: i) apresenta um conteúdo regional dialetal dos falantes do PA São Francisco oriundos de outros estados brasileiros, como Rondônia, Piauí, São Paulo, Pernambuco, Mato Grosso, Mato Grosso do Sul, Espírito Santo, Maranhão e Santa Catarina; e ii) tem a finalidade de ser um *corpus* de estudo, visto que apresenta uma parte da variedade linguística presente no contexto amazônico.

Este *corpus* pode ser representativo devido à diversidade geográfica dos sujeitos-agentes, pois reflete o encontro da diversidade cultural e linguística presente na região amazônica. Os entrevistados, com base em sua vivência no PA São Francisco, realizam comparações entre os seus respectivos estados de origem e as características específicas observadas no projeto de assentamento. Isso reflete elementos importantes de suas experiências de vida na Amazônia e também possibilitam uma análise mais aprofundada da linguagem usada pelos sujeitos-agentes.

Apesar de ser classificado como um *corpus* pequeno (Sardinha, 2004), ele ainda oferece uma quantidade significativa de dados, sendo essa extensão suficiente para permitir análises linguísticas relevantes, como a seleção lexical realizada com o auxílio do programa *Sketch Engine*. O uso desse programa possibilita a compreensão não apenas do conteúdo das entrevistas, mas também das *nuances* e padrões linguísticos presentes no *corpus*, permitindo uma melhor exploração da variedade da língua apresentada na região amazônica.

Portanto, considerando essas características, o *corpus* oral do PA São Francisco pode ser considerado representativo do contexto amazônico, pois não se limita apenas a seleção de elementos linguísticos da língua, como também oferece *insights* sobre a cultura e experiências dos sujeitos-agentes neste contexto. A partir disso, foi possível observar as padronizações e/ou regularidades para indicar as

características desses sujeitos-agentes e também do PA São Francisco.

Sketch Engine: o processo de seleção lexical do *corpus* oral do PA São Francisco

A escolha pelo programa *Sketch Engine*[13] no processo de análise do *corpus* oral do PA São Francisco se justifica pela sua eficácia em explorar e compreender particularidades linguísticas presentes nos dados coletados. Isso significa que o programa é capaz de reconhecer unidades lexicais e contextualizá-las no *corpus*, permitindo uma análise sobre as palavras utilizadas, seus significados e como são empregadas em diferentes contextos temáticos pelos sujeitos-agentes da pesquisa.

Para isso, o programa oferece ferramentas para a exploração de padrões linguísticos e - quando se relaciona com critérios de análise estipulados pelo pesquisador, como, por exemplo, a faixa etária - pode-se compreender como certas palavras ou expressões são utilizadas, ou seja, será possível identificar as diferenças no vocabulário, observar os neologismos, gírias, empréstimos linguísticos e mudanças no significado das palavras. Ele também facilita a manipulação e a análise dos dados transcritos, permitindo a leitura e processamento adequado para a identificação de unidades lexicais relevantes. Além disso, permite ao pesquisador criar o seu próprio *corpus* de estudo e, posteriormente, compará-lo a outros *corpora*, que são disponibilizados pelo *Sketch Engine*.

Pensando nas funcionalidades do programa, optamos por mostrar como se faz a comparação entre um *corpus* de estudo e um *corpus* de referência, para fins de análise linguística. O primeiro refere-se ao *corpus* oral do PA São Francisco, constituído por meio de narrativas dos sujeitos-agentes; enquanto o segundo é o Português Web 2011 (pt TenTen11), adotado para fins de comparação e extração de palavras-chave do *corpus* de estudo.

O *corpus* de referência selecionado está disponível no programa *Sketch*

[13] https://www.sketchengine.eu/. Segundo as informações encontradas no site, o *Sketch Engine* é um dos programas que permite analisar e explorar *corpora* de textos autênticos das línguas e há 600 *corpora* prontos para serem utilizados em mais de 90 idiomas, fornecendo uma amostra representativa de uma língua (Sketch Engine, [2023]).

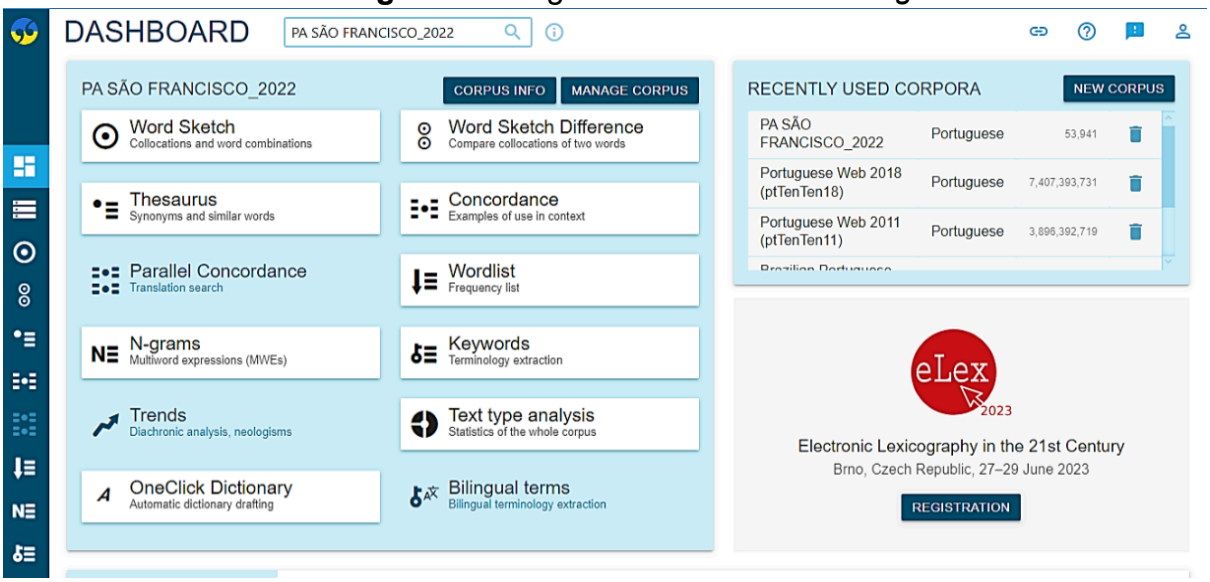
Engine. É um *corpus* de língua portuguesa constituído por textos recolhidos na internet e apresenta quase 4 bilhões de palavras (*tokens),* com as variedades linguísticas do português europeu e do português brasileiro. Comparar os *corpora* foi uma abordagem pensada para verificar se, de fato, o processamento dos dados e a seleção das unidades lexicais realizada pelo programa podem destacar as particularidades, semelhanças e/ou diferenças de um *corpus* de estudo, considerando, nesse caso, a língua portuguesa e as variações linguísticas regionais (diferenças vocabulares, fonéticas, sintáticas e semânticas). Essa análise comparativa também pode revelar as influências culturais e históricas nas formas de expressão linguística de ambos os contextos, visando compreender como a história e a cultura se relacionam com a língua.

Ademais, a comparação pode identificar os padrões lexicais e semânticos particulares do *corpus* oral do PA São Francisco, fornecendo uma compreensão aprofundada das palavras e expressões utilizadas em diferentes contextos culturais. Comparar o *corpus* de estudo com um *corpus* de referência mais amplo, como o Português Web TenTen11, pode ajudar a validar os resultados obtidos no *corpus* de estudo. Isso contribui para evitar generalizações inadequadas, ou seja, não se considera somente a subjetividade do pesquisador, mas também um processamento de dados pelo viés estatístico, proporcionando uma base mais sólida para as considerações sobre a variedade da língua apresentada no contexto amazônico.

Como resultado, foram selecionadas 1.000 palavras individuais (*single-words*) e 1.000 palavras agrupadas (*multi-words terms*), totalizando 2.000 mil palavras. Nesse processo foi possível verificar as unidades lexicais mais frequentes e menos frequentes do *corpus*, possibilitando a organização dos campos lexicais. Contudo, a comparação entre os *corpora* apresentou um número significativo de palavras a serem analisadas, por isso, buscamos as classes gramaticais mais abrangentes do *corpus*. Nesse caso, os substantivos, verbos e adjetivos, que totalizaram 351 unidades lexicais distribuídas entre as três classes.

A aplicação e o funcionamento das ferramentas utilizadas durante o processo de análise do *corpus* de estudo no *Sketch Engine* foram: *Concordance, Wordlist* e *Keywords* (Figura 1).

Figura 1 – Página inicial do *Sketch Engine*

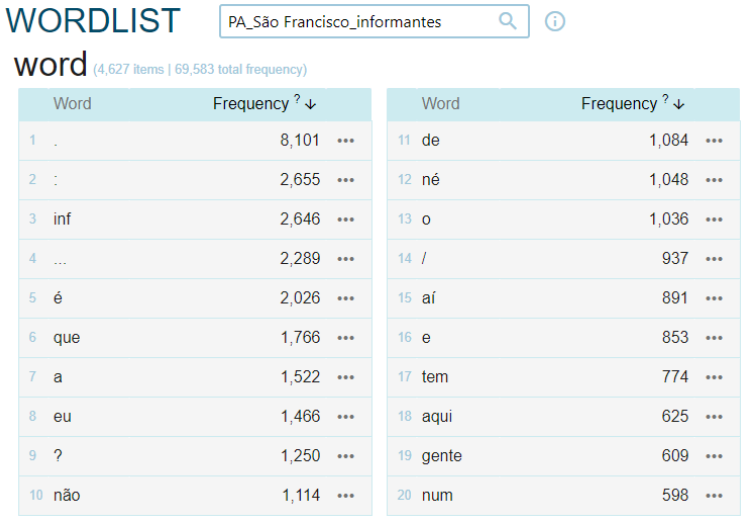


Fonte: Tela do Programa *Sketch Engine.*

A ferramenta *Concordance*, de forma geral, permite encontrar exemplos de uso no contexto, ou seja, é usada para encontrar exemplos de uma palavra em particular junto ao contexto em que ela aparece. Enquanto a *Keywords* permite a localização de palavras-chave que ocorrem com maior ou menor frequência dentro do *corpus* de estudo, o que permite definir ou compreender o tema principal do *corpus*. Por fim, a *Wordlist é* utilizada para gerar uma lista de frequências das palavras, junto aos percentuais.

Das três ferramentas, a *Wordlist* foi a primeira a ser utilizada, tendo em vista que a partir da listagem geral de palavras e suas respectivas frequências, seria possível manipular melhor as outras duas. A ordem das informações aparece da seguinte forma: palavra (*Word)*, frequência *(frequency)* e DOCF, em que mostra quantos documentos diferentes contêm o item. Além disso, a lista é organizada por ordem de frequência e não por ordem alfabética, conforme a Figura 2.

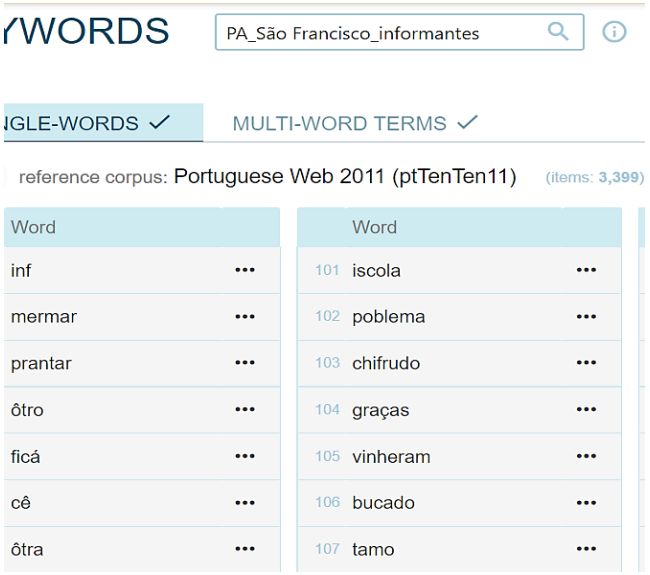
Figura 2 – Amostra da lista por frequência



Fonte: Tela do programa *Sketch Engine.*

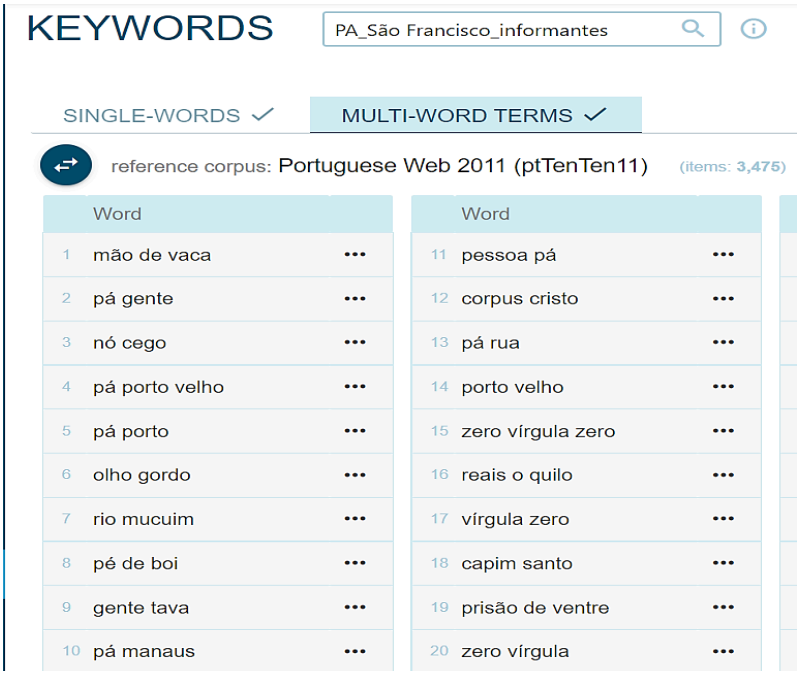
Um ponto observado durante a geração da lista de palavras é de que quando se considera somente o *corpus* de estudo, a opção de listagem é apresentada individualmente como *single-words.* Por outro lado, quando se faz a comparação entre o *corpus* de estudo e o *corpus* de referência, a lista é apresentada tanto como *single-words* quanto em *multi-words terms* (agrupada a duas ou mais palavras), conforme as Figuras 3 e 4.

Figura 3 – Lista de palavras individuais (*Single-words*)



Fonte: Tela do programa *Sketch Engine.*

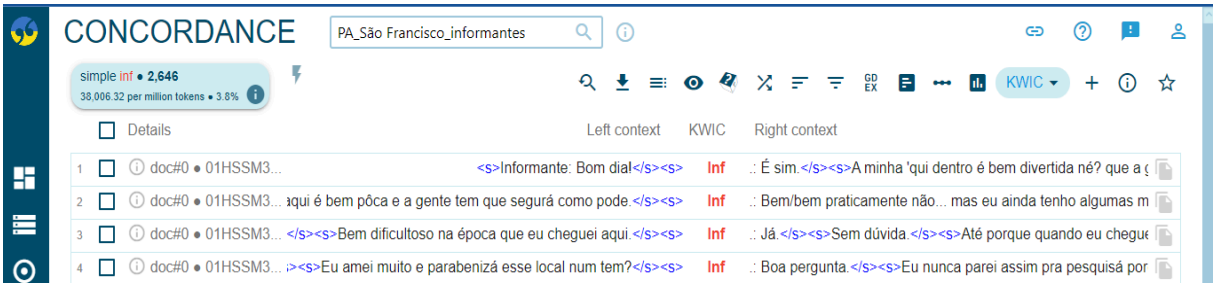
Figura 4 – Lista de palavras agrupadas (*Multi-word terms*)



Fonte: Tela do programa *Sketch Engine.*

Diferente do que ocorre quando se analisa somente o *corpus* de estudo, o programa ao realizar a comparação de dois *corpora* não apresenta de forma explícita a frequência, sendo necessário clicar na pontuação "..." para verificar o número de ocorrências a partir da ferramenta *concordance*. Esta, no que lhe concerne, apresenta o *CQL* para fornecer informações sobre critérios de pesquisa e contagem de frequência (incluindo frequência por milhão e porcentagem de todo o *corpus*). O *Details* mostra em linhas as informações sobre os arquivos. Há também o *left context,* palavra (*token)* localizada à esquerda, a palavra-chave no contexto (*KWIC)* e o *Rigth context*, palavra (*token*) localizada à direita. A Figura 5 mostra como as informações são apresentadas pelo programa.

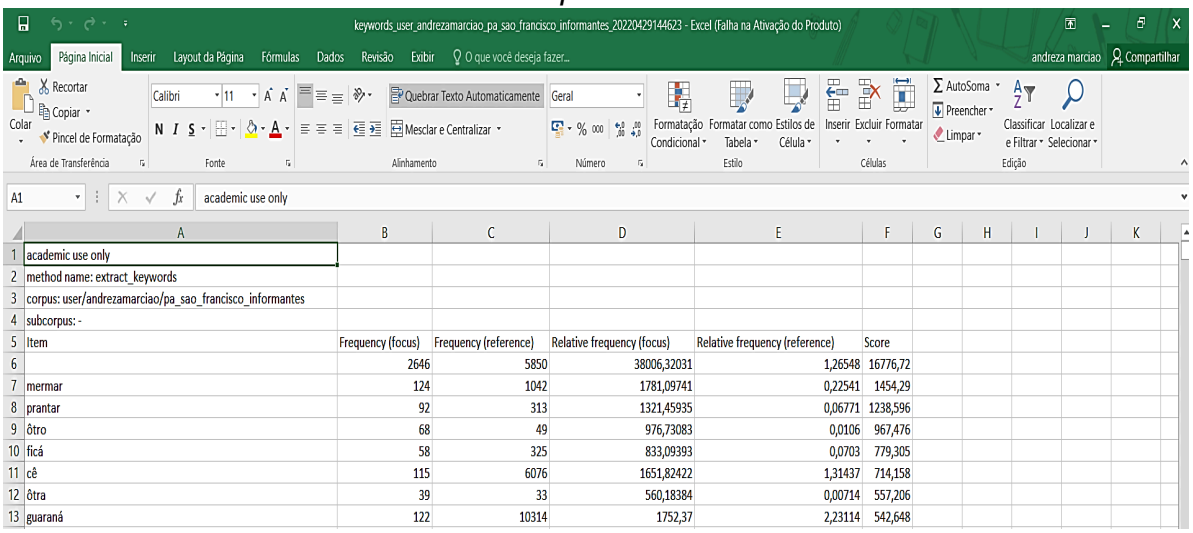
Figura 5 – Ferramenta *Concordance*



Fonte: Tela do programa *Sketch Engine.*

No processo de comparação entre o *corpus* de estudo e o *corpus* de referência, o próprio programa seleciona as palavras-chave (*Keywords*), considerando as frequências que são estaticamente diferentes (maiores ou menores) do que as frequências das mesmas palavras no *corpus* de referência. Sendo assim, é produzida uma lista com as diferenças mais significativas de palavras-chave dentro do *corpus* de estudo, conforme a Figura 6.

Figura 6 – Lista de palavras-chave gerada na comparação entre o *corpus* de estudo e o *corpus* de referência



	A	B	C	D	E	F
1	academic use only					
2	method name: extract_keywords					
3	corpus: user/andrezamarciao/pa_sao_francisco_informantes					
4	subcorpus: -					
5	Item	Frequency (focus)	Frequency (reference)	Relative frequency (focus)	Relative frequency (reference)	Score
6		2646	5850	38006,32031	1,26548	16776,72
7	mermar	124	1042	1781,09741	0,22541	1454,29
8	prantar	92	313	1321,45935	0,06771	1238,596
9	ôtro	68	49	976,73083	0,0106	967,476
10	ficá	58	325	833,09393	0,0703	779,305
11	cê	115	6076	1651,82422	1,31437	714,158
12	ôtra	39	33	560,18384	0,00714	557,206
13	guaraná	122	10314	1752,37	2,23114	542,648

Fonte: Programa *Sketch Engine.*

Com a utilização do *Sketch Engine* e também das ferramentas disponíveis pelo programa foi possível verificar as unidades lexicais que foram mais importantes (mais frequentes) e quais foram as secundárias (menos frequentes) em uma infinita possibilidade de combinações lexicais. Logo, a organização e as análises que

envolveram os campos lexicais se encontram fundamentadas nas informações extraídas do uso real do léxico em diferentes *corpora*.

Segundo Sardinha (2009, p. 91), as palavras-chave podem ser classificadas em dois tipos: “As positivas são aquelas cuja frequência no corpus de estudo é maior do que no corpus de referência, ao passo que as negativas são aquelas cuja frequência é menor no corpus de estudo”. Para efeitos de seleção de lexias, até o momento, as palavras-chave negativas podem ser mais úteis do que as positivas para um estudo, pois, se percebe, por exemplo, que as lexias *prantar* e *guaraná* apresentam um nível de significância relevante para o campo lexical do *trabalho*. Chega-se a essa interpretação devido à seleção das palavras-chave selecionadas dentro do *corpus* de estudo apresentar diferença significativa (maior ou menor) em relação ao *corpus* de referência. Para Sardinha (2009, p. 91)

> O cálculo do que chamamos de ‘maior’ e ‘menor’ é feito pelo programa por meio de testes estatísticos como o qui-quadrado e o log-likelihood, que comparam a frequência de cada palavra no corpus com sua frequência no corpus de referência (caso a palavra em questão não exista no corpus de referência sua frequência será zero).

O pesquisador deve estar atento a leitura realizada pelo programa, principalmente, quando se tratar de um *corpus* oral, pois, sabe-se que a modalidade oral da língua apresenta muitas variações e pode ocorrer que, ao comparar dois *corpora*, tais variações não sejam encontradas em ambos. Isso pode gerar a problemática de análise das formas tipicamente dialetais como *tisôra* e *tesôra*, no *corpus* de estudo, e *tesoura*, no *corpus* de referência. Por isso, uma alternativa seria confirmar manualmente nos *corpora* a ocorrência das variações.

Após a etapa de seleção das palavras-chave geradas de forma semiautomática pelo programa *Sketch Engine*, cabe ao pesquisador considerar os objetivos da pesquisa para selecionar as palavras-chave e analisá-las (Sardinha, 2009). Por exemplo, no Quadro 5, destacam-se alguns fenômenos de variação linguística selecionados pelo programa no *corpus* oral do PA São Francisco, considerando as diferentes origens dos sujeitos-agentes.

Quadro 5 – Variação das lexias e naturalidade dos assentados

Lexias	Tipo de variação	Naturalidade dos assentados
1. a) dificuldade ~ b) dificulidade	Fonético-fonológico	a) PI/MT/ES/RO/PE/MT/MA/SC b) MS
2. a) igarapé ~ b) garapé	Fonético-fonológico	a) MT/RO/PE/MS/PI/SC/SP/ES b) ES/RO/PE/SP/MT
3. a) estilingue ~ b) istilingue	Fonético-fonológico	a) RO/MT/SP/SC/PI/MA/MS b) MS/PE/ES/MT/MA
4. a) baladêra ~ b) baladeira	Fonético-fonológico	a) RO/PI/MA b) RO

Fonte: elaborado pelas autoras.

Com esses dados, foi possível perceber que as lexias selecionadas pelo *Sketch Engine* são representativas do ponto de vista social e cultural dos sujeitos-agentes. No caso de *dificuldade* e *dificulidade,* exemplos apresentados em 1(a) e 1(b), houve referência ao período de acesso ao PA São Francisco durante o inverno amazônico (entre os meses de dezembro e maio), sendo praticamente intrafegável nessa época do ano. Em 2(a) e 2(b), os sujeitos-agentes citam os igarapés existentes no PA como fonte essencial para consumo próprio e cultivo de culturas, bem como servem de orientação para a localização de residências. Os exemplos 3(a), 3(b), 4(a) e 4(b) fazem referência ao brinquedo feito de uma forquilha e duas tiras de borracha, muito presente na infância da maioria dos sujeitos-agentes.

A partir do Quadro 5, foi possível verificar a dinamicidade do léxico e de que ele transpassa barreiras geográficas. O léxico é aberto e infinito, e os assentados do PA São Francisco apresentam uma variação entre si no uso das lexias. No entanto, essa variação não impede ou interfere no processo comunicativo (intercompreensão) desses falantes. Segundo Villalva e Silvestre (2014, p. 23):

> Procurar conhecer o léxico de uma língua a partir do conhecimento do léxico dos falantes implica compreender o que se passa nessa dimensão. O léxico de cada falante, que é também chamado de léxico mental, depende da sua apropriação dos estímulos lexicais a que é exposto, e, portanto, variará muito em função da sua experiência linguística individual, do que ouve, do que lê, do que fala e do que escreve. Um indivíduo não é falante de uma dada língua porque nasceu e cresceu no país onde essa é a língua oficial, mas porque esses foram os dados linguísticos a que foi exposto, enquanto membro de uma dada comunidade, crucialmente nos seus primeiros anos.

O uso variado das lexias pelos assentados representa um incentivo à

ampliação do seu próprio repertório lexical individual, resultando em uma rede de conhecimentos construída e atualizada de acordo com o processo comunicativo coletivo. Além disso, mobiliza os conhecimentos prévios que eles já possuem sobre a língua e o léxico.

Desse modo, afirmamos que o uso da língua pelos sujeitos-agentes do PA São Francisco e os fenômenos linguísticos encontrados não ocorrem de forma aleatória, visto que reflete a sua experiência linguística individual, que é moldada por uma variedade de fatores linguísticos, sociais e culturais. Por isso, o pesquisador deve examinar os dados em busca dos possíveis fatores que influenciam um determinado padrão ou variedade da língua. Portanto, embora o *Sketch Engine* auxilie na leitura e na seleção lexical de um *corpus* oral, acreditamos que os caminhos para uma análise linguística com o auxílio de programas computacionais devem envolver outras questões, como: diferentes abordagens teóricas, consideração dos aspectos linguísticos e extralinguísticos da língua, a experiência do pesquisador com a língua e/ou variedade a ser estudada e também do objetivo de cada pesquisa.

Considerações finais

A partir do aporte metodológico da LC e do auxílio do programa *Sketch Engine* foi possível descrever os processos de processamento e seleção lexical do *corpus* oral do PA São Francisco, considerando um estudo comparativo entre o *corpus* de estudo e o *corpus* de referência, o Português Web 2011 (pt TenTen11).

Nesse percurso, também houve a possibilidade de conhecer a importância das ferramentas do *Sketch Engine*, como a *Concordance, Wordlist* e *Keywords*, na identificação de palavras-chave e na contextualização do léxico encontrado no *corpus* de estudo. Ressalta-se que a seleção de palavras-chave não ocorreu de forma aleatória, mas baseada em diferenças estaticamente significativas de frequência entre os dois *corpora* e chamando a atenção para as variações dialetais na análise de um *corpus* oral.

Com isso, propõe-se que a LC e o *Sketch Engine* oferecem um caminho para formas de análise acerca dos estudos lexicais, bem como outros níveis de análise

linguística, incluindo o léxico e a semântica. A utilização desse aparato e as ferramentas computacionais facilitam a identificação dos padrões e/ou tendências de uma língua, ou uma variedade linguística, mas, cabe sempre ao pesquisador, saber utilizá-las e aplicá-las de acordo com cada objetivo de pesquisa linguística.

Dessa forma, espera-se que essa visão abrangente sobre o processo de constituição e processamento de dados para análise de um *corpus* de estudo contribua para o enriquecimento da compreensão do léxico e das possibilidades de uso da língua realizada por diferentes sujeitos-agentes em diferentes contextos geográficos e culturais.

Referências

AZEVEDO, O.S. *Aspectos dialetais do português da região Norte do Brasil: um estudo sobre as vogais pretônicas e sobre o léxico no Baixo Amazonas (PA) e no Médio Solimões (AM)*. Tese (Doutorado em Linguística) – Centro de Comunicações e Expressão, UFSC, 2013.

BATISTA, B.C.L.L. *Aspectos dialetais do Médio Amazonas: um estudo sobre o léxico. 2019.* Dissertação de mestrado – Universidade Federal do Amazonas, Manaus, AM, 2019.

BENKE, V. C. M. *Tabus linguísticos nas capitais do Brasil*: um estudo baseado em dados geossociolinguísticos. 2012. Dissertação (Mestrado em Estudos de Linguagens) – Universidade Federal de Mato Grosso do Sul, Campo Grande, 2012. Disponível em: https://repositorio.ufms.br/handle/123456789/1836. Acesso em: 17 ago. 2023.

BIBER, D. Representativeness in corpus design. *Literary and Linguistic Computing*, Oxford, v. 8, n. 4, p. 243-257, 1993. Disponível em: https://otipl.philol.msu.ru/media/biber930.pdf. Acesso em: 17 ago. 2023.

BIBER, D. *Variation across speech and writing*. Cambridge: Cambridge University Press, 1998.

CORRÊA, H.C.O. *O falar do caboclo amazonense: aspectos fonéticos-fonológicos e léxico-semânticos de Itacoatiara e Silves*. 1980 – Pontifícia. Universidade Católica do Rio de Janeiro.

CRUZ, M.L.C. *Atlas Linguístico do Amazonas (ALAM)*. 2004. Tese (Doutorado em Letras Vernáculas) – Faculdade de Letras, UFRJ, Rio de Janeiro, 2004.

DAVIES, M. Corpus: an introduction. In: BIBER, D.; REPPEN, R. (ed.). *The*

Cambridge handbook of english corpus linguistics. Cambridge: Cambridge, 2015. p. 11-31.

FILLMORE, C. 'Corpus linguistics' or 'computer corpus linguistics'. *In*: SVARTVIK, J. (org.). *Directions in corpus linguistics*. New York: De Gruyter, 1992. DOI: https://doi.org/10.1515/9783110867275

LAVOR, C. M. A.; ARAÚJO, A. A.; PEREIRA, M. L. S. Os verbos botar e colocar no estado do Maranhão em dados do ALiB: uma pesquisa variacionista. *Leitura*, Maceió, n. 66, p. 146-164, dez. 2020. DOI 10.28998/2317-9945.202066.

LINDQUIST, H.; LEVIN, M. *Corpus linguistics and description of english*. Edinburg: Edinburg University Press, 2009.

MCENERY, A.; HARDIE, A. *Corpus linguistics*. Cambridge: Cambridge University Press, 2012.

PAIM, M. M. T. A variação lexical no campo semântico vestuário e acessórios: um estudo a partir dos dados do Projeto ALiB. *A Cor Das Letras*, Feira de Santana, v. 20, n. 1, p. 204-215, out. 2019. DOI: https://doi.org/10.13102/cl.v20i1.4747

RAYSON, P. Computational tools and methods for corpus compilation and analysis. *In*: BIBER, D.; REPPEN, R. *The Cambridge handbook of english corpus linguistics*. Cambridge: Cambridge, 2015. p. 32-49.

SARDINHA, T. B. *Linguística de corpus*. Barueri: Monole, 2004.

SARDINHA, T. B. Linguística de corpus: histórico e problemática. *D.E.L.T.A.*, São Paulo, v. 16, n. 2, p. 323-367, 2000. DOI: https://doi.org/10.1590/S0102-44502000000200005

SARDINHA, T. B. Questões metodológicas de análise de metáfora na perspectiva da linguística de corpus. *Gragoatá*, Niterói, n. 26, p. 81-102, 2009. Disponível em: https://periodicos.uff.br/gragoata/article/view/33125. Acesso em: 2 fev. 2024.

SARDINHA, T. B.; ALENCAR, A. L. S.; SILVA, C. S.; GIL, C. B.; LOPES, M. J. F.; HUGHES, S. A. S. #eunaovoutomarvacina: uma abordagem da linguística de corpus e da análise multimodal imagética. Intercâmbio, São Paulo, v. 51, p. 178-209, dez. 2022. Disponível em: https://revistas.pucsp.br/index.php/intercambio/article/view/58515. Acesso em: 2 fev. 2024.

SKETCH ENGINE. What is sketch engine?. *Lexical Computing CZ*, Brno, [2023]. Disponível em: https://www.sketchengine.eu/. Acesso em: 17 ago. 2023.
VILLALVA, A.; SILVESTRE, J. P. *Introdução ao estudo do léxico*: descrição e análise do Português. Petrópolis: Vozes, 2014.

Recebido em: 15 dez. 2023.
Aprovado em: 12 fev. 2024.
Publicado em: 30 jun. 2024.

Revisor de língua portuguesa: Juliano Brambilla Neri
Revisor de língua inglesa: Juliano Brambilla Neri
Revisora de língua espanhola: Laura Marques Sobrinho