UNIVERSIDADE FEDERAL DE SANTA MARIA
CAMPUS SANTA MARIA
CENTRO DE ARTES E LETRAS
CURSO DE GRADUAÇÃO BACHARELADO EM LETRAS

Leonardo Mello Ragagnin

SELEÇÃO MACROESTRUTURAL DAS UNIDADES LEXICAIS DO CENÁRIO SOLO PARA O DICIONÁRIO ENCICLOPÉDICO DA AGROECOLOGIA

Santa Maria, RS
Junho, 2023.

SELEÇÃO MACROESTRUTURAL DAS UNIDADES LEXICAIS DO CENÁRIO SOLO PARA O DICIONÁRIO ENCICLOPÉDICO DA AGROECOLOGIA

Trabalho de Conclusão de Curso, apresentado ao Curso de Bacharelado em Letras, da Universidade Federal de Santa Maria (UFSM, RS) – Campus Santa Maria, como requisito parcial para obtenção do título de **Bacharel em Letras.**

Orientador: Profa. Dra. Ana Flávia Souto de Oliveira

Co-orientador: Vitor Jochims Schneider

Santa Maria, RS
Junho, 2023.

Leonardo Mello Ragagnin

SELEÇÃO MACROESTRUTURAL DAS UNIDADES LEXICAIS DO CENÁRIO SOLO PARA O DICIONÁRIO ENCICLOPÉDICO DA AGROECOLOGIA

Trabalho de Conclusão de Curso, apresentado ao Curso de Bacharelado em Letras, da Universidade Federal de Santa Maria (UFSM, RS) – Campus Santa Maria, como requisito parcial para obtenção do título de **Bacharel em Letras.**

Aprovado em 19 de dezembro de 2023:

_____

Ana Flávia Souto de Oliveira
(Orientadora)

_____

Vítor Jochims Schneider
(Coorientador)

_____

Simone Mendonça Soares
(Avaliadora)

Santa Maria, RS
(2023)

DEDICATÓRIA

Dedico este trabalho a todas as pessoas que tenham interesse, queiram saber e conhecer mais sobre a área da Linguística Cognitiva e Lexicografia, e como elas se relacionam entre si e com outras áreas do conhecimento (no caso deste Trabalho de Conclusão de Curso, com a área da Agroecologia).

AGRADECIMENTOS

Agradeço a minha família e aos meus amigos por me apoiarem durante os momentos bons e as dificuldades da vida e vida acadêmica.

À Professora Ana Flávia Souto de Oliveira, por ter sido minha orientadora e ter desempenhado tal função com dedicação e amizade.

Ao Professor Vítor Jochims Schneider, por ter sido meu co-orientador e ter desempenhado tal função com dedicação e amizade.

Aos demais professores, pelas correções e ensinamentos que me permitiram apresentar um melhor desempenho no meu processo de formação profissional e acadêmica ao longo do curso.

Às pessoas com quem eu convivi, ao longo desses anos de curso na UFSM, na bolsa do CCSH, na bolsa da Central de Periódicos, no PET Letras- Laboratório *Corpus* e na Empresa Grámmatos Júnior, que me incentivaram e que certamente tiveram impacto na minha formação profissional e acadêmica.

"A verdadeira dificuldade não está em aceitar ideias novas, mas escapar das antigas."

(John Maynard Keynes)

RESUMO

SELEÇÃO MACROESTRUTURAL DAS UNIDADES LEXICAIS DO CENÁRIO SOLO PARA O DICIONÁRIO ENCICLOPÉDICO DA AGROECOLOGIA

AUTOR: Leonardo Mello Ragagnin
ORIENTADOR: Profª Drª Ana Flávia Souto de Oliveira

Este trabalho tem como objetivo apresentar e discutir questões relacionadas à metodologia de seleção das unidades lexicais da Agroecologia para a construção de um dicionário enciclopédico dessa área. O projeto leva em consideração os aportes da Lexicografia e da Linguística Cognitiva. Portanto, este estudo tem como referenciais teóricos, no que diz respeito à construção de um projeto lexicográfico e conceitos da lexicografia, os trabalhos de Atkins e Rundell (2008); Oliveira (2010); Oliveira; Gatti, Pipper (2022). Além disso, a obra de Fillmore (1982) serve como referência acerca da Semântica de *Frames*, pois uma palavra ou expressão evoca um *frame* e para compreender uma unidade lexical, deve-se ter acesso à estrutura na qual ela se encaixa. Com isso, a análise e discussão das unidades lexicais do cenário SOLO se baseiam nos *frames*. Por fim, o dicionário leva em consideração as necessidades do público-alvo, que tem como finalidade principal a aquisição de conhecimentos no âmbito da Agroecologia no que se refere ao aumento da matéria orgânica e micro-organismos, que mantém o solo vivo.

Palavras-chave: Agroecologia; Dicionário enciclopédico; Solo; Lexicografia; Linguística Cognitiva.

ABSTRACT

MACROSTRUCUTRAL SELECTION OF LEXICALS ITEMS OF THE SOIL SCENARIO FOR THE ENCYCLOPEDIC DICTIONARY

AUTHOR: Leonardo Mello Ragagnin
ADVISOR: Ana Flávia Souto de Oliveira.

This work aims to present and discuss questions related to methodology of selections of the lexicals items of Agorecology for building of the dicitionary in this area. This Project takes into consideration the contributions of Lexicgraphy and Cognitive Linguistics. Thus, this study has as theoretical basis regarding the construction of lexicographic projects and concepts of Lexicography, based on the work of Atkins and Rundell (2008); Oliveira (2010); Oliveira, Gatti and Pipper(2022). In addition, Fillmore's work(1982) serves as reference, Frames Semantics, because a word or expression. Evokes a frame and to understand a lexical unit, one must have acess to the lexical structure in wich it fits. Hence, the analysis and discussion of the lexical items of the solo scenario are based on the frames. Finally, the dictionary takes into account the needs of the lay public and the students from agricutural schools, its main finality is to acquire knowledges within the Agorecology scope concernig: increasing organic matter and microorganisms, which maintain the soil alive.

Keywords: Agroecology; Encyclopaedic Dictionary; Soil; Lexicography; Cognitive Linguistics.

LISTA DE FIGURAS

Figura 1 – AntConc 1. 33
Figura 2 – Frequência AntConc. 34
Figura 3 – Resultado de busca da tupla *macroagregados, bioquímica do solo, solo vivo e agroecologia* nas páginas da internet. 36
Figura 4 – Resultado de busca da tupla *bioquímica do solo, macroagregados e solo vivo e agroecologia* nas páginas Web . 38
Figura 5 – Resultado de busca da tupla *proteção do solo, manejo de solo e fertilidade do solo* nas páginas Web . 38

LISTA DE QUADROS

Lista 1 – Sementes do primeiro *corpus*. 35
Lista 2 – Sementes do segundo *corpus*. 36

LISTA DE TABELAS

Tabela 1 –Resultados dos termos e expressões na *Word List e Cluster* 40

SUMÁRIO

1. INTRODUÇÃO 13
2. OBJETIVOS 16
2.1. OBJETIVO GERAL 16
2.2. OBJETIVOS ESPECÍFICOS 16
3. JUSTIFICATIVA 16
4. REFERENCIAL TEÓRICO 18
4.1.LINGUÍSTICA COGNITIVA 18
4.2.SEMÂNTICA DE *FRAMES* 19
4.3.LEXICOGRAFIA 21
4.4.LEXICOGRAFIA COGNITIVA 24
4.5.*CORPUS* 29
5. METODOLOGIA 34
6. RESULTADOS APRESENTADOS DO *CORPUS* 39
7.CONSIDERAÇÕES FINAIS 41
REFERÊNCIAS 42

1. INTRODUÇÃO

Os dicionários são obras de referências e repertórios linguísticos que apresentam os significados das palavras socialmente determinados por uma comunidade linguística. Além disso, essas obras atribuem o valor de verdade e conhecimento não só sobre a língua e seus usos, mas também sobre os diversos campos do conhecimento. Segundo Atkins e Rundell (2008, p.2),

> um dicionário é uma descrição do vocabulário usado por membros de uma comunidade de fala (por exemplo, por "falantes de inglês"). E o ponto de partida para essa descrição é a evidência do que os membros da comunidade de fala fazem quando se comunicam entre si.

Por outro lado, o *corpus* é um conjunto de dados linguísticos coletados, em grande volume, tanto nos textos escritos quanto nos falados, que podem ser compilados manualmente ou com ferramentas computacionais de *corpora* (utilizado como plural de *corpus).* Por isso, o uso de *corpus* é muito importante para a Lexicografia, pois permite ao lexicógrafo extrair grandes quantidades de itens lexicais presentes nos textos de um determinado domínio (área temática). Os registros desses itens lexicais podem ser relevantes não só para constituir a macroestrutura de algum dicionário, mas também para obter diversas informações sobre o uso desses itens.

Além disso a macroestrutura, segundo Bugueños (2006), é o conjunto de entradas de um dicionário e que, para as expressões linguísticas fazerem parte da macroestrutura do dicionário, é necessário levar em conta os tipos de usuários, de dicionários e os objetivos desejados; utilizando-se, portanto, de critérios de seleção macroestrutural quantitativa, qualitativa e de lematização das palavras. Sobre a macroestrutura, ela será discutida nos próximos tópicos.

Os registros do *corpus* de cada dicionário têm suas peculiaridades, pois necessitam ser representativos do domínio específico. Com isso, neste trabalho, são utilizados como base teórica os preceitos da Semântica Cognitiva, particularmente da Semântica de *Frames,* para interpret compreensão às unidades lexicais[1] do Cenário[2] SOLO do domínio específico da Agroecologia. Isso porque não é possível

[1] As unidades lexicais, segundo Fillmore(1982), são definidas como os significados dos lemas, isto é, o pareamento de uma palavra e um sentido dela constitui uma unidade lexical.

[2] "Unidade de conhecimento que serve como base para o entendimento da expressão" (Oliveira, Piper, Gatti, 2022, p.85).

interpretar ou compreender uma palavra ou expressão de forma isolada, ou seja, o léxico é apresentado a partir do domínio da Agroecologia e do Cenário SOLO, e os *frames* que dão sustentação para o entendimento das palavras e expressões evocadas por este cenário.

Neste sentido, o presente trabalho se insere na temática a partir da pesquisa *Semântica Cognitiva e aplicações web*, que tem como finalidade apresentar os métodos de análise e seleção das unidades lexicais as quais entrarão na macroestrutura do dicionário, pertencente ao cenário SOLO do domínio da Agroecologia. Inicialmente, o intuito da pesquisa era construir um Caderno de Campo para produtores rurais, para isso, foram realizadas algumas entrevistas em propriedades rurais e nas escolas agrícolas, e também, durante o estudo, foi feita uma pesquisa sobre os dicionários da área de Produção Rural disponíveis no mercado e notou-se que há mais dicionários voltados para especialistas na área, em outros termos, para fins acadêmicos. Percebendo que havia uma lacuna de dicionários voltados para alunos de escolas agrícolas e para o público leigo, a orientadora e os integrantes do projeto de pesquisa *Semântica Cognitiva e aplicações web* optaram por compilar *corpora* e planejar a construção de um Dicionário Enciclopédico direcionado para área da Agroecologia, baseando-se no conhecimento enciclopédico dos sujeitos e tendo como modelos os dicionários enciclopédicos já disponíveis, alicerçados na *FrameNet* e que utilizam a Semântica de *Frames.*

Sobre *a* elaboração do *corpus* do *Dicionário Enciclopédico de Agroecologia*, ela tem como características: uma linguagem não muito técnica, ou seja, as palavras e os conceitos, presentes no contexto dos estudantes e agricultores (sem ensino superior), não devem ser acadêmicos; um público-alvo voltado para estudantes de escolas agrícolas, agricultores familiares (somente os que não têm uma formação acadêmica) e leigos na área; e o Dicionário que será construído com base nesse *corpus* serve para consulta de saberes, principalmente, para a aquisição de novos saberes na área da Agroecologia. Nesse viés, o *Dicionário Enciclopédico da Agroecologia* apresenta suas particularidades tanto do *Corpus* quanto do público. Por esse motivo, para ajudar na extração dos dados linguísticos, será compilado um *corpus,* que busca representar de forma genérica os conhecimentos vinculados à área de Produção Rural.

Com base nisso, este trabalho tem como objetivo apresentar e discutir sobre a metodologia de seleção macroestrutural das unidades lexicais que farão parte do Cenário SOLO. A escolha para trabalhar esse Cenário surgiu pelo fato de o conhecimento sobre solo ser muito importante para a manutenção do equilíbrio vital dos seres vivos, e também para auxiliar os estudantes de escolas agrícolas e o público leigo a manejar o solo corretamente.

Tendo em vista a perspectiva enciclopédica do Dicionário, objetivo e o público alvo, este Trabalho de Conclusão de Curso tem como aportes a Linguística Cognitiva e a Lexicografia. A escolha pela Linguística Cognitiva se dá, porque essa vertente da Linguística é considerada imprescindível para a construção dos conhecimentos lexicais estabelecida pela relação do conhecimento enciclopédico (experiência e conhecimento de mundo) com a mente e o uso da linguagem. Com base na Linguística Cognitiva, foi utilizado como referencial teórico a Semântica de *Frames*. No que se refere aos *frames*, eles são estrutura de conceitos que, para conseguir compreender as unidades lexicais, estas precisam estar associadas com a estrutura na qual elas se encaixam (Fillmore,1982), isto é, para conseguir compreender e interpretar expressões como *biomassa seca*, *qualidade do solo* etc, estas precisam ativar conhecimentos relacionados ao cenário SOLO (que evocam cenas da experiência humana sobre recursos naturais e seus usos, e sobre as práticas da Produção Rural, os quais são armazenadas na mente dos indivíduos). Na perspectiva da Lexicografia, este trabalho focará na metodologia empregada para a Seleção Macroestrutural das expressões e termos, coletados e analisados nos *corpora*, relacionados ao Cenário SOLO que entrarão na macroestrutura do *Dicionário Enciclopédico da Agroecologia.*

Ademais, este trabalho está dividido em três partes. Na primeira seção, serão apresentados os objetivos do trabalho e a justificativa alinhada aos referenciais teóricos como: Atkins e Rundell (2008); Chishman (2016); Oliveira(2010); Oliveira, Pipper e Gatti (2022); Oliveira e Echevarria (no prelo); Fillmore (1982); Evans e Green (2006); Biderman (2001); Ferrari (2011); Welcker (2004); Moreira Filho (2021); Martelotta e Palomanes (2021); e Kader e Richter (2013).

Na próxima seção, será apresentada como recorte metodológico a compilação de um *corpus* obtido por meio do *software BootCat* e sua análise com o *AntConc*.

Na seção seguinte, serão expostos os resultados apresentados com relação aos dados obtidos no *corpus*, objetivando que este projeto lexicográfico, a partir da definição do público alvo e da sua finalidade, torne-se parte da descrição do Cenário Solo presente em um *Dicionário Enciclopédico da Agroecologia*, elucidando a importância da área da Agroecologia.

2. OBJETIVOS

2.1. OBJETIVO GERAL

O objetivo deste trabalho é apresentar e discutir questões metodológicas relacionadas à seleção de unidades lexicais do cenário solo na Agroecologia para a construção de um dicionário dessa área a partir da construção de um *corpus* específico desse domínio.

2.2. OBJETIVOS ESPECÍFICOS

- Definir critérios para a seleção das unidades lexicais do cenário solo a partir do público alvo e de suas necessidades;
- Analisar o cenário solo a partir da Semântica de *Frames*.

3. JUSTIFICATIVA

Este trabalho se justifica por apresentar às metodologia utilizadas para a realização deste trabalho: como a *Semântica de Frames* e a aplicação *web* presentes nos dicionários já elaborados serão utilizadas como modelo para construção do *Dicionário Enciclopédico da Agroecologia*, bem como o uso de ferramentas computacionais para fazer a Análise e a Seleção das unidades lexicais relacionadas ao cenário SOLO, o qual futuramente fará parte da construção do projeto lexicográfico, *Dicionário Enciclopédico de Agroecologia*, que surgiu no projeto de pesquisa de *Semântica Cognitiva e aplicações web*, quando os integrantes do grupo estavam realizando um Caderno de Campo e as entrevistas com produtores rurais da região de Santa Maria. Com isso foi percebido que muitos dicionários das áreas das Produção Rural são mais voltados para o público acadêmico. Além disso, esse Dicionário visa auxiliar na consulta dos saberes e na aquisição de conhecimentos tanto dos estudantes de escolas agrícolas quanto do público leigo.

A escolha para trabalhar com a análise e seleção das unidades lexicais do Cenário SOLO surgiu devido a relevância que o solo tem para a manutenção de vida das plantas, e também por ele estar interligado aos outros cenários como: Adubação, Consorciação de Culturas, Controle de Pragas entre outros.

O Dicionário de Agroecologia é Enciclopédico, pois leva em consideração o conhecimento de mundo e o contexto dos sujeitos. Segundo Ferrari (2011), esse conhecimento enciclopédico, baseado no contexto, na cognição e no uso da linguagem, define a construção de significados das palavras ou expressões, e também aponta que o conhecimento não é desorganizado, isto é, ele é estruturado em rede. Dessa forma, são apresentadas as unidades lexicais do cenário SOLO a partir do domínio *Agroecologia*, e os *frames* que dão sustentação para a compreensão e entendimento dessas unidades. Embora cada Cenário tenha suas palavras e expressões que são específicas, eles se relacionam entre si. Por isso, o conhecimento enciclopédico dos sujeitos, como os dos agricultores e estudantes de escolas agrícolas, e também a escolha dessa base teórica da Linguística Cognitiva, são fundamentais para a construção de acepções das palavras e termos, relacionados aos seus respectivos Cenários, que entrarão na Macroestrutura do Dicionário.

A pesquisa lexicográfica baseada em *corpora* é muito importante para a Lexicografia, pois permite obter um grande volume de dados linguísticos compilados para visualizá-los e manipulá-los, beneficiando a elaboração de diversos projetos lexicográficos. A partir do *corpus,* é possível verificar uma diversidade de textos em que os itens lexicais dele estão inseridos, sendo recomendável para os dicionários. Segundo os autores Oliveira, Hatwig e Pipper (2021) e Moreira Filho (2021), com a chegada da Internet, o uso de *corpus* foi sendo muito disseminado, tornando-se imprescindível para o trabalho dos lexicógrafos e linguistas e facilitando as tarefas desses profissionais, por meio do uso de ferramentas computacionais, na busca rápida de grandes quantidades de dados linguísticos listados, contabilizados e comparados, pois se esse processo de análise for realizado por analistas humanos, em outras palavras, feito manualmente, isso se torna muito exaustivo e demorado (Moreira Filho, 2021). Corroborando com o ponto de vista do autor Moreira Filho (2021):

> Determinadas tarefas não são indicadas para um analisador/ pesquisador humano, por exemplo, a listagem, contagem e comparação de grandes quantidades de palavras. Os programas de análise de *corpora* podem realizar tarefas repetitivas em grandes quantidades de dados com maior consistência e abrangência. Tais programas podem contar milhares de palavras em diversos textos, extrair palavras-chave a partir de comparação e cálculo de fórmulas estatísticas, além de gerar visualizações privilegiadas dos contextos de palavras de busca (Moreira Filho, 2021, p.26).

Na próxima seção, Referencial Teórico, serão tratados, esses aspectos abordados neste tópico como: as definições de Linguística Cognitiva, Semântica de *Frames*, a Lexicografia e Lexicografia Cognitiva.

4. Referencial Teórico:

4.1. Linguística Cognitiva

A Linguística Cognitiva, surgida entre o final da década de 1970 e o início da década de 1980, é uma vertente da Linguística que surgiu da insatisfação dos autores Fillmore, Langacker, Lakoff, Talmy e Fauconnier com a modularidade da mente, proposta pela Linguística Gerativa de Noam Chomsky, a qual distinguia os módulos relacionados a linguagem dos demais módulos associados ao raciocínio matemático, espacial, percepção entre outros(Ferrari, 2011; Martelotta; Palomanes, 2021). Além disso, esses autores também contestaram a forma como os linguistas formais diferenciavam o conhecimento linguístico do não linguístico, em outras palavras, havia uma distinção entre a Semântica como conhecimento linguístico e a Pragmática enquanto conhecimento não linguístico, voltado para o contexto e o uso da linguagem. Os linguistas formais concentravam a atenção em estudar mais sobre o conhecimento linguístico do que o conhecimento pragmático ou não linguístico e, em outros termos, a Pragmática era considerada uma área periférica para os Estudos Linguísticos (Evans; Green, 2006).

Na perspectiva da Linguística Formal, o conhecimento linguístico é equivalente ao conhecimento de dicionário (Semântica), e o conhecimento extralinguístico é o conhecimento enciclopédico (Pragmática). Contudo, não há distinção do que é Semântica e Pragmática para os linguistas cognitivos, pois essas áreas formam um *continuum* na perspectiva da Linguística Cognitiva, mesmo que a Linguística Cognitiva não desconsidere a existência da área da Pragmática. Corroborando com esse ponto de vista, os autores Evans e Green (2006) afirmam que a Semântica Cognitiva não separa a visão enciclopédica da visão de dicionário e, embora o significado das palavras esteja armazenado na memória de longo prazo, o significado das palavras muda de acordo com o contexto, ou seja, o uso que os falantes de uma comunidade linguísticas fazem da língua, e é isso o que determina o significado dos termos e expressões.

Ainda segundo esses mesmos autores, a construção dos significados dos itens lexicais é definida por meio do conhecimento enciclopédico, o qual não é caótico, e sim organizado e construído em rede, e ela é dinâmica e de natureza conceitual. Além disso, a centralidade está associada à saliência de alguns aspectos do conhecimento enciclopédico, relacionados a um termo, são importantes para o significado deste.

Os estudiosos Fillmore e Langacker propuseram as suas teorias com base no conhecimento enciclopédico. Charles Fillmore propôs a Semântica de *Frames*. Langacker, a noção de domínios. As teorias desses autores determinam que as palavras ou construções gramaticais só podem ser compreendidas e interpretadas com a estruturas de conceitos vinculados aos *frames* ou aos domínios associados a elas. Por exemplo, para entender o conceito da palavra *biomassa* esta deve estar associada ao *frame* SOLO, do domínio da Agroecologia.

4.2. Semântica de *Frames*

A Semântica de *Frames* foi criada pelo Linguista da Universidade da Califórnia e do Instituto Internacional de Ciências da Computação e líder do Projeto *FrameNet*, Charles Fillmore. Esse autor se fundamentou na teoria de Psicologia de Gestalt, por meio dos conceitos de Figura e Fundo, para distinguir o conceito lexical específico, denominado de figura, da estrutura do todo a qual esse conceito específico está relacionado, definida como fundo. Por exemplo, é impossível entender o que é uma hipotenusa (conceito específico) sem a relação dela com o *frame* (estrutura do todo) Triângulo Retângulo. Com isso, a abordagem da Semântica de *Frames* foi proposta, baseando-se nessa teoria da Psicologia e na contestação da modularidade da mente, proposta pela Linguística Gerativa (Evans; Green, 2006), em outras palavras, Fillmore só concordava com o único pressuposto dessa teoria de Chomsky que é “a linguagem é o espelho da mente” (Ferrari, 2011, p. 13).

Segundo Evans e Green (2006), a Semântica de *Frames* surge dentro da Semântica Cognitiva como uma tentativa de desvendar as propriedades de um inventário de conhecimento estruturado associado com as palavras e considera quais as consequências das propriedades desse sistema de conhecimento que podem servir para um modelo semântico. Os *frames* relacionam elementos e

entidades com uma cena particular culturalmente incorporada da experiência humana, os quais são representados em nível conceitual e mantidos na memória de longo prazo. Portanto, os *frames* são estruturas de conceitos que para compreender uma palavra ou expressão, estas devem estar associadas a esses *frames* e as cenas da experiência humana. Por exemplo: para compreender e interpretar a palavra *manejo de solo*, ela deve estar relacionada ao cenário SOLO que evoca cenas da experiência da humana ligadas às práticas sustentáveis de nutrir, corrigir e cultivar o solo.

A Semântica de *Frames* leva em conta a visão de mundo dos indivíduos, como também os estilos e os registros de usos da língua tanto na fala quanto na escrita desses sujeitos. Como exemplo disso: a palavra *bicicleta* tem a definição de "veículo de duas rodas" se estiver ligada ao domínio de veículos; e também tem o significado de "forma de chute em que o jogador de pernas para cima chuta a bola diretamente para o gol, como se estivesse pedalando uma bicicleta", se estiver associado ao domínio Futebol. Por outro lado, a Valência de um *frame* é uma forma em que os itens lexicais podem ser combinados com outras palavras para formar sentenças, ou seja, a valência está relacionada aos números de participantes e argumentos de um verbo ou item lexical. Além do mais, segundo as informações do site da *FrameNet*, Fillmore utiliza outros conceitos para construir a Semântica de *Frames* como: Elemento de *frame*, que seria o papel semântico de um *frame* específico que é a unidade básica do *frame;* Herança em que estabelece a relação do *frame* filho como um tipo específico do *frame* pai (principal); *frame* semântico é uma estrutura descritiva para caracterizar o significado lexical. Esses conceitos da Semântica de *Frames* são utilizados pelo projeto da *FrameNet*, o qual será descrito no subitem de Lexicografia Cognitiva, para descrever o léxico dos Dicionários já elaborados por meio de uma aplicação de lexicografia computacional.

4.3. Lexicografia

Conforme Welker (2004), a Lexicografia se divide em duas partes: a Lexicografia Prática que está associada a ciência, técnica ou prática de criar dicionários; a Lexicografia Teórica, denominada também de Metalexicografia, estuda sobre Tipologia, crítica relacionada aos dicionários, pesquisa histórica sobre a Lexicografia e elaboração dos dicionários.

Para pensar a escrita de um Dicionário, os autores Atkins e Rundell (2008) afirmam que, ao se planejar a sua escrita, deve ser levado em conta que o seu espaço é limitado, pois não há como inserir em dicionário todos itens lexicais que existem tanto nas línguas naturais como nas diversas áreas de conhecimentos. Por isso, o dicionário é sempre incompleto, ou seja, a elaboração deste é um trabalho em progresso. O dicionário é uma ferramenta para registro do léxico (vocábulo ou nomeação de entidades) de uma determinada língua natural. Segundo Biderman (2001), os conceitos ou significados de um determinado léxico são formas de registrar os fenômenos percebidos por meio das experiências sensoriais e da cognição. Nesse registro, o dicionário pode ter conceitos de palavras que dizem respeito às classes gramaticais, ortografias, pronúncias e traduções (se for um dicionário bilíngue ou trilíngue) dessas, assim como pode ter o significado de palavras e expressões de uma determinada área do saber, pois todos esses registros presentes no dicionário são baseados no contexto, nas experiências e na cognição dos indivíduos que estão inseridos em uma determinada comunidade linguística.

Para planejar a escrita de um projeto lexicográfico, os pressupostos dos autores Atkins e Rundell (2008) devem ser levados em consideração como: mercado de dicionários, tipo de dicionário; a linguagem empregada nele; o número de entradas na macroestrutura e o espaço; o formato dele, se será impresso, eletrônico ou com uma aplicação web; se o perfil do dicionário é leigos, estudantes de escolas, estudantes acadêmicos ou especialistas em uma determinada área; como também o conhecimento enciclopédico de uma determinada comunidade linguística deve ser levado em conta para a construção do significados das palavras. Ainda conforme esses autores, os linguistas e os lexicógrafos não precisam ter o domínio de todas as áreas (Atkins; Rundell, 2008), porém para trabalhar com a escrita de um dicionário os campos relevantes para isso são: Semântica Lexical, a Teoria Cognitiva e a Pragmática. Esses domínios auxiliam em melhorias na escrita de dicionários.

No que diz respeito à macroestrutura, Segundo Rey-Bove *apud* Oliveira (2010), a macroestrutura é um conjunto de entradas ordenadas. Por outro lado, de acordo com Bejoint (2000) *apud* Oliveira (2010), a Macroestrutura é definida como conjunto de entradas em que a seleção da macroestrutura é baseada no critério de seleção por frequência. Apoiando-se nessa perspectiva, Welker (2004, p.96) *apud* Oliveira (2010) propõe que o principal padrão que determina a inclusão da

macroestrutura é a frequência fornecida por meio das pesquisas em *corpora.* Ainda sobre a frequência, Burgueño (2007a, p.265) *apud* Oliveira, afirma que uma unidade lexical que tenha uma frequência menor do que número mínimo estabelecido não deve ser incluída na macroestrutura. Segundo Atkins e Rundell (2008), a Macroestrutura é responsável pela organização dos tipos de entradas incluídos no dicionário e como a lista de palavras-chave são organizadas. Ainda conforme esses mesmos autores, para organizar as entradas e a lista de palavras-chave que farão parte da macroestrutura do dicionário, alguns fatores devem ser levados em conta como: os tipos de usuários, o tamanho do dicionário, o custo de produção do dicionário, tipos de vocabulários, domínio, a frequência de um item lexical em um corpus, a relevância do item lexical e o quão familiar um item lexical é para os usuários do dicionário (Atkins; Rundell, 2008).

Devem ser levados em consideração os conceitos de lema, unidade lexical e item lexical ao trabalhar na decisão dos tipos de entradas e como organizar a lista de palavras-chave para compor a macroestrutura do Dicionário. Primeiramente, a palavra lema é utilizada para designar as palavras-chave e todas as suas formas. Já as unidades lexicais formam os blocos de construção das entradas da macroestrutura, ou seja, esses blocos dizem respeito à constituição da macroestrutura de um dicionário. Os sentidos atribuídos a uma unidade lexical são enumerados, os quais são válidos somente para essa unidade lexical. O item lexical é qualquer palavra, abreviação, palavra parcial ou expressões multipalavras que podem figurar no dicionário (muitas vezes como a palavra-chave da entrada), como objeto de descrição lexicográfica, geralmente é uma definição ou tradução (Atkins; Rundell, 2008).

Com isso, é a partir do reconhecimento do item lexical, por meio de uma palavra, combinações de palavras, expressões multipalavras ou frase presentes nos dados de um *corpus*, que poderá ser tratado como as entradas constituintes da macroestrutura de um Dicionário. Sobre as expressões multipalavras e as frases fixas e semifixas presentes no *corpus*, o lexicógrafo deve estar atento a elas, e também precisa saber diferenciá-las durante a análise do *corpus* antes de incluí-las como entradas no Dicionário, levando em conta a semântica, os contextos e os exemplos de uso delas em uma sintaxe. A seleção dos tipos de vocabulários que receberão *status* de palavra-chave deve ser decidida conforme os domínios (área temática do dicionário, como Produção Rural, Linguística, Ciências Cognitivas,

Ciências Computacionais etc), o estilo (como as expressões ou como a linguagem são utilizadas de acordo com a área, se são poéticas, burocráticas, jurídicas etc), registro (formal ou informal) entre outros fatores. Alguns critérios devem ser adotados para selecionar os itens lexicais, como: a frequência do item no *corpus*; a relevância do item; o quão conhecido o item é para os usuários do dicionário; se há possíveis traduções e suas conotações ou significados adicionais (Atkins; Rundell, 2008).

Após essa análise desses itens lexicais no *corpus*, para organizar a lista de palavras-chave alguns critérios devem ser considerados como: alfabetização, silabificação e homógrafos. A alfabetização é a ordem alfabética de como os lemas (palavras-chave em todas suas formas) aparecerão na macroestrutura, se eles serão apresentados na forma de letra por letra ou palavra por palavra. Silabificação é a marcação de sílabas das palavras ou uma quebra de palavras. As palavras-chave homógrafas são duas ou mais palavras escritas de forma semelhante em que cada uma recebe um número e são tratadas como entidades diferentes (Atkins; Rundell, 2008).

Tendo esse planejamento em vista, a construção do Dicionário Enciclopédico será elaborada por meio de uma aplicação *web,* tendo como perfil de usuário o público leigo e os estudantes de escolas agrícolas, a linguagem presente nele não é muito técnica e há poucos dicionários de Produção Rural destinados para estudantes escolares e público leigo, no mercado. Sobre o que diz respeito ao cenário SOLO do Dicionário, o número de entradas (conjunto de palavras e expressões ou unidades lexicais) na macroestrutura será definido com relação ao *frame* SOLO, e também com base no corpus de referência (os documentos do MAPA, Embrapa e Emater) e no conhecimento enciclopédico dos alunos de escolas agrícolas e do público leigo. Baseado também nas concepções de Macroestrutura, o projeto do Dicionário Enciclopédico, principalmente no que se refere a Cenário SOLO, utiliza os critérios de frequência, domínio, relevância dos termos e como os termos são conhecidos pelos usuários do dicionário.

4.4. Lexicografia Cognitiva

A Lexicografia Cognitiva se baseia na Linguística Cognitiva, tendo como contribuição teórica a Semântica de *Frames* proposta por Charles Fillmore. Segundo

Atkins e Rundell (2008), a Semântica de *Frames* é muito relevante para o trabalho dos lexicógrafos, pois essa teoria de Fillmore descreve as palavras, seus diversos significados e como essas são combinadas para formar enunciados e sentenças em uma determinada língua natural. Conforme esses autores, a abordagem dos *frames semânticos,* para a descrição do comportamento das palavras, é mais útil para trabalhar com os itens lexicais presentes em um *corpus*, assegurando que nenhum dado do *corpus* seja esquecido, e também esses *frames* são relevantes para o estudo da língua e da Lexicografia, pois podemos nos comunicar por meio de uma língua porque as palavras e frases as quais usamos para evocar suas estruturas em nossas mentes, para que possamos compartilhar uma interpretação do que é dito ou escrito.

Essa teoria foi utilizada no Projeto *FrameNet*. Esse projeto foi criado por Fillmore, na década de 1990, o qual propôs uma aplicação de lexicografia computacional, baseada na Semântica de *Frames,* na qual esta mostra as palavras em Inglês, seus sentidos e como elas são combinadas para formar frases. Nesse projeto, segundo as informações do site do Instituto Internacional de Ciências da Computação (*International Computer Science Institute/ICSI),* as palavras são agrupadas de acordo com os frames semânticos; tipos de situações de representações esquematizadas (como *comer, remover* etc), das quais eles participam; os participantes; e os padrões nos quais eles se combinam com outras palavras e frases próximas a eles são descritas de acordo com a maneira como os *elementos de frames* são expressos.

Sobre a *FrameNet,* as autoras Oliveira e Echevarria (no prelo) a definem como:

> um recurso lexical que descreve os *frames* da língua inglesa que dão suporte ao significado de expressões linguísticas, com base em evidências de *corpus* e na descrição sintático-semântica (incorporando representação das valências dessas expressões). Como exemplo, o *frame* **Cure** [*cura*] é definido da seguinte forma: “Esse *frame* lida com um Curador que trata e cura uma Condição (as feridas, a doença ou a dor) de um Paciente, muitas vezes também mencionando o uso de um Tratamento ou Medicação particular [...]” (sv. *Cure*, FrameNet). A base de dados apresenta unidades lexicais que evocam o cenário e exemplos anotados de sentenças relacionadas ao frame, além da descrição de seus elementos centrais e não centrais e dos tipos de relação que o cenário descrito mantém com outros cenários (Oliveira; Echevarria, no prelo, p.8-9).

Como base nessas informações, muitos dos Dicionários que circulam na Internet utilizam como referência a Semântica de Frames presente na *FrameNet* como: o *Dicionário da Copa do Mundo* 2014, desenvolvido pelo Laboratório *FrameNet* Brasil da Universidade de Juiz de Fora (UFJF); *Dicionário Field*, desenvolvido pelo grupo de pesquisa da Universidade do Vale do Rio dos Sinos (UNISINOS); o *Dicionário Olímpico*, do grupo de pesquisa da Universidade do Vale do Rio dos Sinos (UNISINOS); e o *Lexicovid-19*, do grupo de pesquisa de Lexicografia e Linguística Cognitiva da Universidade Federal de Santa Maria (UFSM), *projeto Lexicovid-19*.

O primeiro Dicionário foi influenciado pelo modelo de construção lexical da *FrameNet* e surgiu com a finalidade de auxiliar os turistas que vieram para o Brasil, durante a Copa do Mundo de 2014 a conhecer as palavras e expressões que evocam cenas relacionadas aos domínios Turismo e Futebol, e as cenas que são evocadas por essas palavras e expressões. Ao pensar nesses domínios associados ao Brasil, os indivíduos pensam nas cenas da experiência humana que são associadas aos pontos turísticos do Brasil, à competição entre países e à Copa de 2014. Além das unidades lexicais e das cenas, o dicionário oferece frases em que essas aparecem. Este projeto da *FrameNet* Brasil foi elaborado com parcerias da *FrameNet* (rede de lexicografia computacional) e a *International Computer Science Institute* (dos Estados Unidos). Por outro lado, o Dicionário é ofertado em três línguas: Português, Inglês e o Espanhol; o usuário pode buscar por uma palavra, digitar frases ou ver somente o significado de cada palavra ou expressão. Ao clicar no Menu e em uma palavra, o aplicativo exibe a cena em que ela está vinculada; a valência dela, como os participantes, lugar, meio etc; e as sentenças em ela aparece. Por exemplo: a busca pela palavra *bandeirinha* exibe como informações:

- Cena: a arbitragem;
- Definição: árbitro auxiliar de uma partida de futebol;
- Traduções para o inglês, *linesman* e *assistant referee;*
- Tradução para o Espanhol: *assistente*;
- Participantes: Função, papel por cada membro da arbitragem no jogo do futebol; Especificação, especificação da função desempenhada pelo membro da arbitragem; Descrição, alguma característica do árbitro que não diga respeito à sua função na partida;

- Mais palavras (os *frames* filhos vinculados ao *frame* pai *bandeirinha)*: arbitragem, árbitro, auxiliar, juiz, quarto árbitro e trio de arbitragem.

O segundo Dicionário foi criado pelo grupo SemanTec da UNISINOS, utilizando como inspiração a FrameNet, que se baseia na Semântica de Frames, para elaborar o *Dicionário Field*. A proposta do grupo da UNISINOS em elaborar o *Dicionário Field* de forma online foi semelhante ao do grupo FrameNet Brasil que era criar um Dicionário voltado para os domínios do futebol e da Copa, o qual seguia o mesmo molde da *FrameNet*. O Dicionário é oferecido em três línguas: Português, Inglês e Espanhol, por meio de uma aplicação *web;* nela, o usuário pode buscar por Palavras ou Cenários. Ao contrário do Dicionário da Copa do Mundo que usa a cena (como noção de frame, *frame* semântico ou unidade básica para entendimento dos itens lexicais), o *Dicionário Field* utiliza o Cenário para definição de *frame* semântico. Ao buscar pela expressão *jogo de corpo,* o site exibe como resultados da busca:

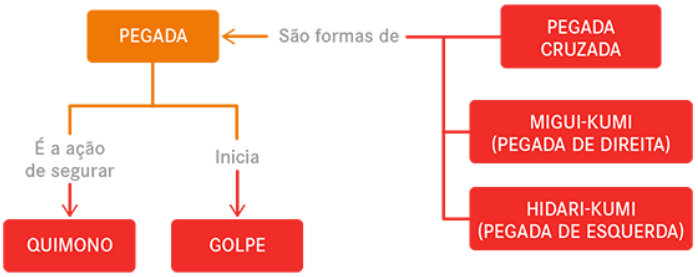
- Cenário: Marcação;
- Sinônimo em Português: **chega-pra-lá**
- Frases em Português: Damião foi para cima da zaga e ganhou a bola no jogo de corpo; Na tentativa de fazer um **jogo de corpo**, o atacante perdeu a bola; Neymar leva **chega-pra-lá** e devolve com tapinha na cabeça de zagueirão;
- Traduções para o Inglês: ***body check****;*
- Sinônimo em Inglês: **shoulder charge;**
- Frases em Inglês: Arne Friedrich was fortunate to stay on the pitch after a cynical ***body check*** on Rooney; Kompany saw yellow for a **body check** on Hazard; A **shoulder charge** is most commonly used by a defender to knock a player in possession and win back the ball;
- Definição com Tradução em Espanhol: En el contexto del fútbol brasileño, “jogo de corpo” es una representación figurativa que simboliza utilizar el cuerpo como una herramienta, o sea, el jugador se utiliza de su cuerpo para alejar un rival del balón, o para ganar una pelota dividida, por ejemplo;
- Palavras relacionadas ao frame *Jogo de Corpo:* cercar, fechar a marcação e marcação homem a homem.

Após a elaboração do *Dicionário Field* em 2014, o grupo *SemanTec* decidiu criar um Dicionário, com aplicação *web* e baseado na *FrameNet*, *o Dicionário*

Olímpico. Esse dicionário apresenta duas formas de busca como lista de itens lexicais e lista de cenários (*frames* semânticos), e também oferece exemplos de usos em sentenças. Ao contrário dos dicionários já citados, este dicionário é bilíngue, tendo tradução dos termos e das frases para o Inglês. O usuário, ao clicar na opção de Esporte *Judô* e no Cenário *Pegada,* recebe como resultados:

- Cenários relacionados: realizada por Judocas, Fazer parte da Luta e Vem antes de Técnicas de Projeção;
- Palavras relacionadas: pegada, pegada cruzada e pegada de esquerda;

Mapa do *frame Pegada*



Fonte: Site do Dicionário Olímpico

O quarto Dicionário, o Lexicovid-19, conforme Oliveira e Echevarria, apresenta:

> As categorias informacionais, como definições, infográficos, notícias, artigos e vídeos explicativos, são apresentadas ao consulente a partir de três microestruturas: uma para o domínio, uma para o cenário e uma para a unidade lexical. Antes de visualizar o significado do item lexical, são disponibilizadas, tanto nos cenários quanto nos domínios, diferentes tipos informações referentes ao contexto em que esse item está inserido, atribuindo maior contextualização ao significado das unidades lexicais e facilitando a sua compreensão (Oliveira; Echevarria, no prelo, p.11).

Ainda, segundo essas autoras, os Cenários, os Domínios e as Unidades Lexicais do dicionário aparecem no site como:

> Cada um dos cenários que dependem do domínio acessado para o seu entendimento recebe também uma microestrutura com informações constantes: definição, unidades lexicais, vídeo, Infográfico e Saiba mais. O cenário **Medidas de prevenção**, por exemplo, abarca unidades lexicais relacionadas ao conhecimento sobre intervenções não farmacológicas destinadas a diminuir o contágio do vírus entre as pessoas, como *barreira sanitária, distanciamento social, etiqueta respiratória, lockdown, quarentena e uso de máscara*. Além disso, **Medidas de prevenção** necessita de uma matriz de outras estruturas de conhecimento de caráter mais genérico para sua composição, como SAÚDE, POLÍTICA, NOVO CORONAVÍRUS e ECONOMIA (Oliveira; Echevarria, prelo, p.13).

Portanto, seguindo essa linha desses trabalhos já realizados os quais seguem os pressupostos da Semântica de Frames e têm inspiração na *FrameNet,* o *Dicionário Enciclopédico da Agroecologia* busca uma aproximação com esses dicionários para utilizar a Semântica de *Frames*, a partir dos *Cenários,* que serviram para descrever as unidades lexicais e oferecer uma base de amplo acesso às informações. Sobre o Cenário SOLO, serão seguidas apenas algumas aplicações da *FrameNet* por ser Cenário estático, ou seja, não serão consideradas as valências atribuídas aos lemas e suas unidades lexicais, pois muitas dessas palavras desse Cenário não são verbos, os quais exigem participantes e argumentos vinculados a eles.

Na seção seguinte, será apresentada a metodologia utilizada para extrair, analisar e selecionar as unidades lexicais mais frequentes e representativas do Cenário SOLO, ou seja, como as ferramentas computacionais de extração e análise de *corpora* serão utilizadas para fazer a compilação dos *corpora*, *o corpus 1 e corpus 2,* seja na extração deste, seja na análise quantitativa e qualitativa dos termos relevantes para entrar na Macroestrutura do cenário estudado.

5. METODOLOGIA

O *corpus* é um conjunto de textos falados e/ou escritos em língua natural armazenados de modo disposto e informado, que pode ser lido e digitalizado pelos computadores ou demais dispositivos tecnológicos. Essa coletânea de textos pode ser representações de uma língua ou de uma variedade linguística de alguma comunidade linguística. Antes do surgimento da Internet, compilar *corpora* e manipulá-los em grande volume era muito caro, demorava muito tempo e necessitava de uma grande equipe para trabalhar com *corpora*. Com o surgimento da web e das ferramentas computacionais, o uso dos *corpora* têm sido expandido,

permitindo trabalhar com grandes volumes de *corpora* e com diversos tipos destes, e também tem facilitado o trabalho dos linguistas e lexicógrafos, tornando a tarefa desses profissionais rápida, econômica e eficaz. Sobre a expansão da *web,* o autor Moreira Filho (2021) afirma que:

> Cresce, cada vez mais, a quantidade de *corpora* disponíveis, tendo em vista os benefícios da pesquisa baseada em corpus. Podemos encontrar *corpora* disponíveis na Internet, em diferentes línguas e tipos (gerais, especializados, históricos, escritos, de fala etc.). Muitos estão à disposição gratuitamente, com licença para pesquisa e estudo. Contudo, há limitações para pesquisas e usos mais extensivos, dada a questão de direitos autorais de textos (Moreira Filho, 2021, p.26).

Em compensação, a representatividade dos *corpora* dependerá do tipo de objetivo e da variedade coletada neles, isto é, para estudar e coletar termos e expressões e verificar o tipo de linguagem presentes no domínio da Agroecologia e que circulam no meio do público leigo e dos estudantes de escola agrícola, a compilação dos *corpora* desta área deverá ser representativo para o objetivo desejado. Por outro lado, segundo Leech (1992) *apud* Kader e Richter (2013), para planejar a compilação de um *corpus* e para que esta seja bem sucedida, alguns critérios devem ser seguidos como:

> a) por maior que seja um corpus, ele é apenas uma amostra da língua em uso;
> b) os dados a analisar não devem ser escolhidos de acordo com as preferências do pesquisador, e sim aleatoriamente, e nenhum deles pode ser considerado irrelevante para a pesquisa;
> c) teorias ou modelos podem ser criados para explicar os dados encontrados (a partir da intuição ou experiência do investigador, por exemplo), mas os valores quantitativos do modelo devem ser obtidos dos dados do corpus;
> d) a precisão do modelo pode ser testada em outro corpus;
> e) a princípio,a qualidade de um modelo pode ser medida e comparada com a de outros modelos(essa interação é importante para que os modelos de desempenho linguístico sejam progressivamente aperfeiçoados)e diferentes modelos podem ser testados com o mesmo corpus,de forma que a superioridade de um modelo em relação a outro possa ser demonstrada (Leech 1992 *apud* Kader; Richter, 2013).

A compilação dos *corpora* do projeto leva em consideração esses critérios propostos por Leech (1992), no caso deste trabalho, os itens lexicais relacionados ao cenário SOLO da análise da amostra dos *corpora* devem ser significativos para entrar na macroestrutura do Dicionário; será utilizado um *corpus* de referência baseados nos documentos da Embrapa, Emater e MAPA, como modelo, para fornecer palavras-chave ao programa extrator de *corpus*; a compilação do *corpus* do Cenário SOLO será comparada com outro *corpus* do mesmo Cenário para verificar

quais unidades lexicais aparecem com mais frequência e se um desses *corpora* não tem termos técnicos e/ou ambíguos de acordo com o domínio da Agroecologia e o perfil de usuário do Dicionário.

Sobre as ferramentas de *corpora,* o autor Moreira Filho trata sobre a importância delas para a compilação dos *corpora* e quais são os tipos de ferramentas de *corpora* utilizadas:

> As ferramentas computacionais são geralmente utilizadas para a reorganização e a extração de informações no corpus para observação e interpretação de dados, fornecendo novas perspectivas para a análise linguística. As ferramentas computacionais mais comuns são:
> ● Programas para listar palavras - fazem a contagem das palavras em um corpus;
> ● Concordanciadores - programas que permitem que o usuário procure por palavras específicas em um corpus, fornecendo exaustivas listas para as ocorrências da palavra em contexto;
> ● Etiquetadores - fazem análises automáticas do corpus e inserem etiquetas (códigos) de ordem morfossintática, sintática, semântica ou discursiva (Moreira Filho, 2021, p.26) .

Esses conceitos sobre *corpora* e ferramentas de *corpora*, este trabalho não tem como finalidade tratar sobre a Linguística de *Corpus* e o uso desses softwares com viés da programação ou da Linguística Computacional*,* e sim utilizar os *corpora* e o uso das ferramentas computacionais como metodologias empregadas no processo de elaboração do projeto lexicográfico *Dicionário Enciclopédico de Agroecologia*. Essas ferramentas são utilizadas para extrair grandes quantidades de dados linguísticos, presentes em textos da área da Produção Rural ou Agroecologia, que são significativos para compor a Macroestrutura do Dicionário e entrar nela. Para isso, são levados em conta somente os itens lexicais do Cenário SOLO para trabalhar com a relevância destes para fazer parte da macroestrutura do Dicionário. Reiterando isso, este estudo trata somente da compilação e análise dos *corpora* do Cenário SOLO e não de discutir toda a elaboração do Dicionário. Para a compilação dos *corpora* relacionados ao Cenário SOLO e em relação aos tipos de softwares que Moreira Filho (2021) trata, foram utilizados: Programas para Listar Palavras e Concordanciador, o *AntConc*; enquanto o programa *BootCaT* é usado só para Extração de *Corpus.*

O *BootCaT* é um *Software* utilizado para compilar um *corpus* no qual o usuário seleciona as palavras e expressões de interesse e as fornece ao programa para a coleta de textos. Essa seleção dos itens lexicais ou palavras-chave são

denominadas de sementes (*seeds*). Essas sementes são combinadas em tuplas (*tuples*), as quais servem de suporte para a coleta de textos na web. A partir dos termos e expressões que o usuário forneceu ao programa, são extraídos textos coletados na web, que contêm essas combinatórias de sementes.

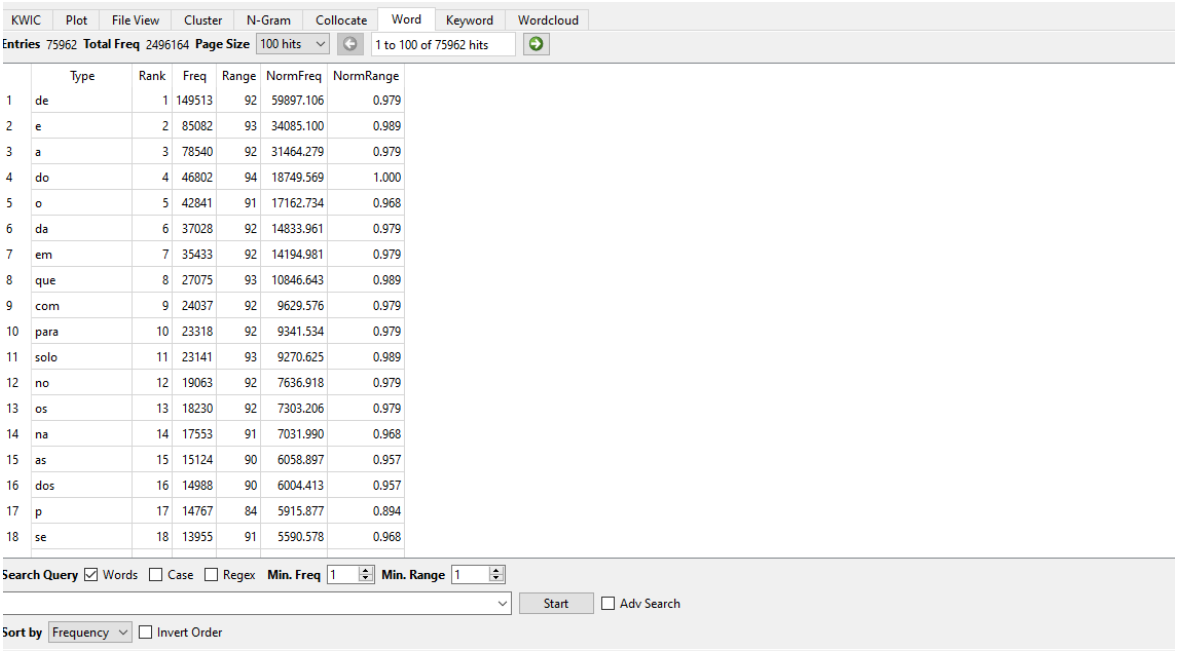
Para que a seleção e construção do *corpus* ocorra, o usuário deve fornecer ao programa, no mínimo, 5 sementes; definir um comprimento de tupla para as combinatórias, de no mínimo, 3 itens lexicais, para extração de textos de páginas web em que essas sementes aparecem; é necessário especificar o site de busca (Google, Bing, Yahoo etc); limitar o número de páginas web que o *software* retorna, para cada tupla, ao usuário; e o usuário pode excluir as combinatórias não relevantes para o *corpus*.

O *AntConc* é um *software* de análise de *corpora* que foi desenvolvido pelo Professor Lawrence Anthony da Faculdade de Ciência e Engenharia da Universidade de Waseda, Japão. Esse programa oferece um guia de ferramentas: *KWIC* (*Keywords in Context* ou Palavras-chave em Contexto), *Cluster*s, *N-grams*, *Collocate* (colocações), *Word List* (lista de palavras) e *KeyWord List* (lista de palavras-chave). Primeiramente, antes de acessar cada guia do *AntConc*, o usuário deve carregar seu *corpus* e abrir para fazer a análise do *Corpus* construído. A funcionalidade *KWIC* permite que o usuário veja as palavras ou expressões inseridas nas frases dos textos coletados no *Corpus*. Essa guia oferece exemplos desses itens lexicais em uma frase ou em um texto, isto é, contextos de usos dos termos ou expressões.

O *Cluster* funciona como um guia de pesquisa em que permite ao usuário buscar por uma palavra ou por um padrão no programa e agrupe os resultados da pesquisa juntamente com as palavras ordenadas à esquerda ou à direita do termo pesquisado. Além disso, esse indivíduo pode estabelecer um número máximo e mínimo do comprimento das palavras no Cluster, a ordenar a forma como os termos ou padrões e dos resultados aparecerão no Cluster e inverter a posição destes. A diferença entre o *Cluster* e *N-grams* é que: o *Cluster* define os padrões de sequência de palavras, enquanto o *N-grams* determina o número de palavras presentes no *corpus*, porém na funcionalidade *N-grams* o usuário não pode dispor do modo como os termos ou padrões aparecerão, isto é, se eles serão apresentados tanto na direita quanto na esquerda do *N-grams*.

A guia *Collocate* permite que o usuário busque as colocações de uma palavra ou expressão. As colocações de uma palavra ou expressão podem ser ordenadas pela frequência total e frequência à direita ou à esquerda do termo pesquisado, como também as colocações podem ser ordenadas por valores estatísticos. Isso é exemplificado por meio da Figura 1:

Figura 1: AntConc

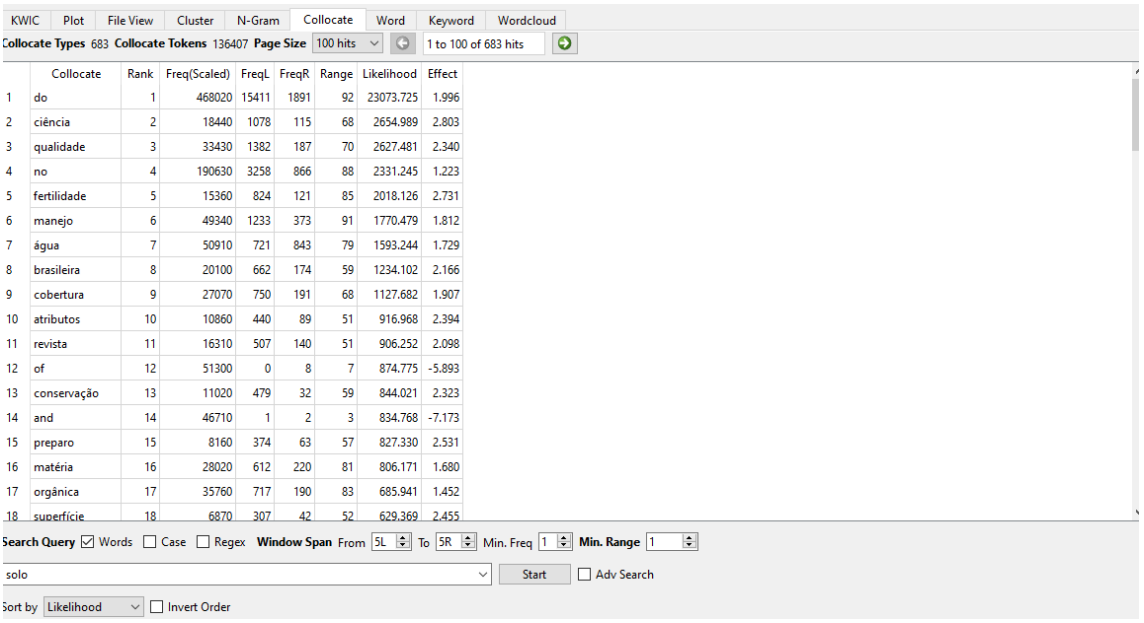


	Type	Rank	Freq	Range	NormFreq	NormRange
1	de	1	149513	92	59897.106	0.979
2	e	2	85082	93	34085.100	0.989
3	a	3	78540	92	31464.279	0.979
4	do	4	46802	94	18749.569	1.000
5	o	5	42841	91	17162.734	0.968
6	da	6	37028	92	14833.961	0.979
7	em	7	35433	92	14194.981	0.979
8	que	8	27075	93	10846.643	0.989
9	com	9	24037	92	9629.576	0.979
10	para	10	23318	92	9341.534	0.979
11	solo	11	23141	93	9270.625	0.989
12	no	12	19063	92	7636.918	0.979
13	os	13	18230	92	7303.206	0.979
14	na	14	17553	91	7031.990	0.968
15	as	15	15124	90	6058.897	0.957
16	dos	16	14988	90	6004.413	0.957
17	p	17	14767	84	5915.877	0.894
18	se	18	13955	91	5590.578	0.968

Fonte: AntConc.

A funcionalidade *Word List* contabiliza o número de palavras ou expressões encontradas no *corpus* e gera uma lista ordenada desses itens lexicais. *Word List* permite que o usuário encontre os termos ou expressões mais frequentes no *corpus*. Os itens lexicais podem ser ordenados pelo número de frequência. Por exemplo, a Figura 2 demonstra isso:

Figura 2: Frequência AntConc



	Collocate	Rank	Freq(Scaled)	FreqL	FreqR	Range	Likelihood	Effect
1	do	1	468020	15411	1891	92	23073.725	1.996
2	ciência	2	18440	1078	115	68	2654.989	2.803
3	qualidade	3	33430	1382	187	70	2627.481	2.340
4	no	4	190630	3258	866	88	2331.245	1.223
5	fertilidade	5	15360	824	121	85	2018.126	2.731
6	manejo	6	49340	1233	373	91	1770.479	1.812
7	água	7	50910	721	843	79	1593.244	1.729
8	brasileira	8	20100	662	174	59	1234.102	2.166
9	cobertura	9	27070	750	191	68	1127.682	1.907
10	atributos	10	10860	440	89	51	916.968	2.394
11	revista	11	16310	507	140	51	906.252	2.098
12	of	12	51300	0	8	7	874.775	-5.893
13	conservação	13	11020	479	32	59	844.021	2.323
14	and	14	46710	1	2	3	834.768	-7.173
15	preparo	15	8160	374	63	57	827.330	2.531
16	matéria	16	28020	612	220	81	806.171	1.680
17	orgânica	17	35760	717	190	83	685.941	1.452
18	superfície	18	6870	307	42	52	629.369	2.455

Fonte: AntConc

Ambos os programas podem ser utilizados tanto em computadores que têm o sistema operacional Windows quanto os que têm Linux e MacOs. O *software* *AntConc* pode ser baixado gratuitamente no site do Professor Laurence Anthony (https://www.laurenceanthony.net/*software*/antconc/). Para este trabalho, o *BootCaT* será utilizado para construir o *corpus* e o *AntConc*, por meio das funcionalidades *Cluster*, *N-grams* e *Word List*, para fazer a análise qualitativa e quantitativa, e a seleção das unidades lexicais relevantes, que entrarão na macroestrutura do *Dicionário Enciclopédico da Agroecologia* no Cenário SOLO.

De modo a compilar um *corpus*, buscaram-se termos e expressões relevantes para o Cenário SOLO. A pesquisa desses dados linguísticos se baseia na relevância que essas expressões apresentam nos textos científicos da Agroecologia sobre solos, nas Fichas Agroecológicas do MAPA e de documentos da Embrapa sobre solos. Esses materiais são relevantes e servem como um *corpus* de referência, pois estes documentos visam difundir a Agroecologia, com as práticas sustentáveis, principalmente as que estão relacionadas ao solo, tanto para os especialistas e estudantes de escolas agrícolas quanto para o público leigo.

Após essa busca, esses dados linguísticos foram inseridos no *software* Extrator de *corpus*, o *BootCaT*. Foram feitas duas tentativas de elaborar um *corpus*. Como primeira tentativa de compilar um *Corpus*, foram fornecidas para o programa

10 sementes relacionadas ao Cenário SOLO, recolhidas do material mencionado acima. Isso é exemplificado por meio da Lista 1, Sementes do primeiro *corpus*:

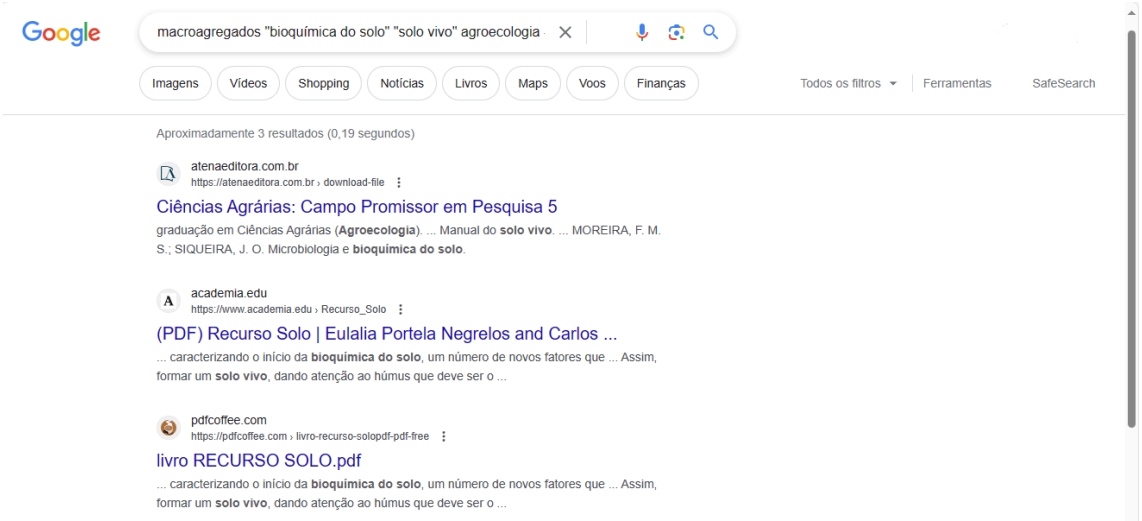
Lista 1: Sementes do primeiro *corpus*

Biomassa
Bioquímica do solo
Calagem corretiva
Fertilidade sistêmica do solo
Macroagregados
Macroporos
Microagregados
Microporos
Proteção do solo
Solo vivo

Após isso, foi definido um comprimento de 4 tuplas por combinatória, o número máximo de 10 páginas web retornadas, inserida a palavra *agroecologia* manualmente em todas as tuplas para que o programa retornasse textos da perspectiva agroecológica e, com isso, foram geradas tuplas com as seguintes combinatórias: *bioquímica do solo, macroagregados, calagem e agroecologia; bioquímica do solo, macroporos, macroagregados e agroecologia; proteção do solo, microporos, solo vivo e agroecologia; solo vivo, bioquímica do solo, calagem e agroecologia; solo vivo, macroporos, biomassa e agroecologia; solo vivo, macroporos, calagem e agroecologia; macroagregados, bioquímica do solo, solo vivo e agroecologia; macroagregados, biomassa, microagregados e agroecologia.* Como site de busca para coletar os links presentes na pesquisa das páginas da *web* (*queries*), foi definido o Google. Definido o site de busca, foram salvas todas as *web queries*. A partir disso, o programa fez a coleta das URLs, por meio das tuplas geradas no *software*, porém não houve tantos resultados relevantes para a busca dos termos inseridos no *software*.

Como exemplo disso, a primeira imagem das tuplas geradas pelo *BootCaT* exibe 3 buscas nas páginas da *Web*:

Figura 3: Resultado de busca da tupla *macroagregados, bioquímica do solo, solo vivo* e *agroecologia* nas páginas da internet.



Fonte: Google

Na segunda tentativa de compilar o *corpus*, devido à limitação das escolhas lexicais, foi repensada e revista a seleção dos itens lexicais inseridos na primeira tentativa de compilação do *corpus*, pois estes eram muito específicos e técnicos para entrar no Cenário SOLO. Com isso, foram inseridas outras palavras e expressões relevantes, precisas e não muito técnicas para o Cenário SOLO. Isso é exemplificado por meio da Lista 2, Sementes do segundo *corpus* 2:

Lista 2: Sementes do segundo *corpus*

Biomassa
Bioquímica do solo
Calagem corretiva
Fertilidade do solo
Manejo convencional
Manejo de solo
Proteção do solo
Qualidade do solo

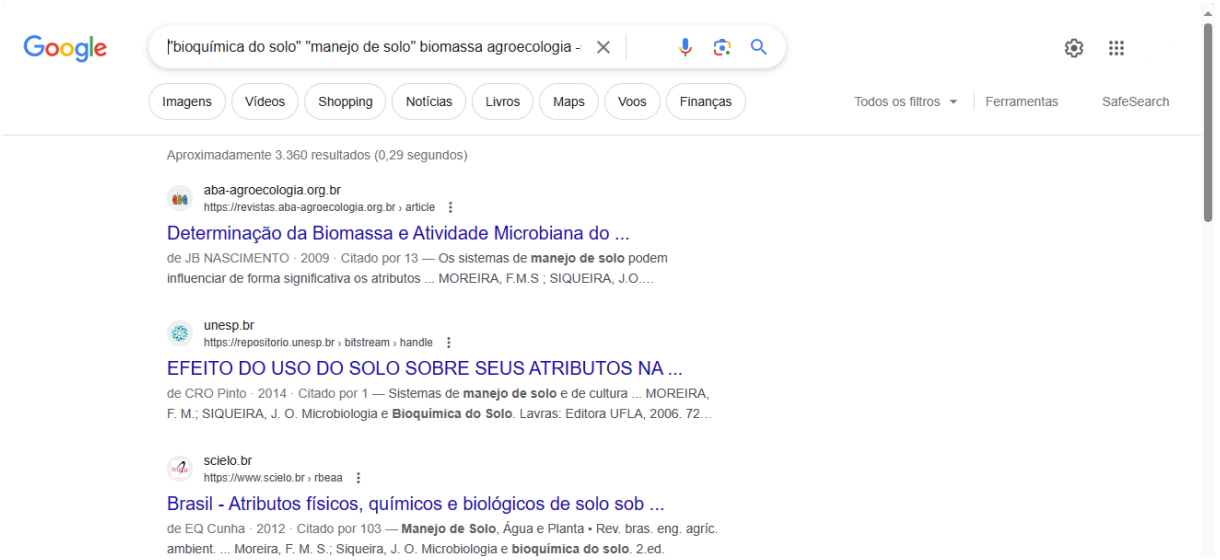
Solo vivo
Agroecologia

Após isso, foi determinado um comprimento de 4 tuplas, o número máximo de 20 páginas web retornadas e foram gerada tuplas com as seguintes combinatórias: *bioquímica do solo, fertilidade do solo, biomassa e agroecologia; bioquímica do solo, manejo do solo, biomassa e agroecologia; bioquímica do solo, proteção do solo, fertilidade do solo e agroecologia; bioquímica do solo, proteção do solo, solo vivo e agroecologia; bioquímica do solo, qualidade do solo, solo vivo e agroecologia; calagem corretiva, manejo convencional, fertilidade do solo e agroecologia; fertilidade do solo, calagem corretiva, qualidade do solo e agroecologia; fertilidade do solo, manejo convencional, calagem corretiva e agroecologia; fertilidade do solo, manejo de solo, e agroecologia; fertilidade do solo, solo vivo, manejo de solo e agroecologia; manejo convencional, manejo de solo, bioquímica do solo e agroecologia; manejo convencional, solo vivo, biomassa e agroecologia; manejo de solo, biomassa, manejo convencional e agroecologia; proteção do solo, calagem corretiva; biomassa e agroecologia; proteção de solo, fertilidade do solo, solo vivo e agroecologia; proteção do solo, manejo do solo, fertilidade do solo e agroecologia; proteção do solo, qualidade do solo, calagem corretiva e agroecologia; qualidade do solo, fertilidade do solo, solo vivo e agroecologia; solo vivo, bioquímica do solo, manejo convencional e agroecologia.* Foi repetido o mesmo procedimento do *corpus* anterior. Como site de busca para coletar os links presentes na pesquisa das páginas da *web* (*queries*), foi definido o Google. Definido o site de busca, foram salvas todas as *web queries* e após isso, foram coletadas as URLs para buscar e selecionar textos em que aparecem as diversas combinações dessas palavras e expressões nos materiais pré-selecionados.

Após a coleta dessas informações no *corpus* obtido no *BootCaT*, verificou-se que o resultado dessas buscas foi promissor, tendo muitos registros desses termos e expressões nas páginas coletadas na web. Para ter bastantes resultados na busca de textos na web em que essas combinações aparecem, foi acrescentada a palavra "agroecologia” para situar o domínio da Agroecologia e facilitar a busca dos termos e expressões selecionados para o Cenário SOLO.

Por exemplo, a segunda imagem das tuplas geradas pelo *BootCaT* apresenta 3.360 resultados coletados nas páginas da web:

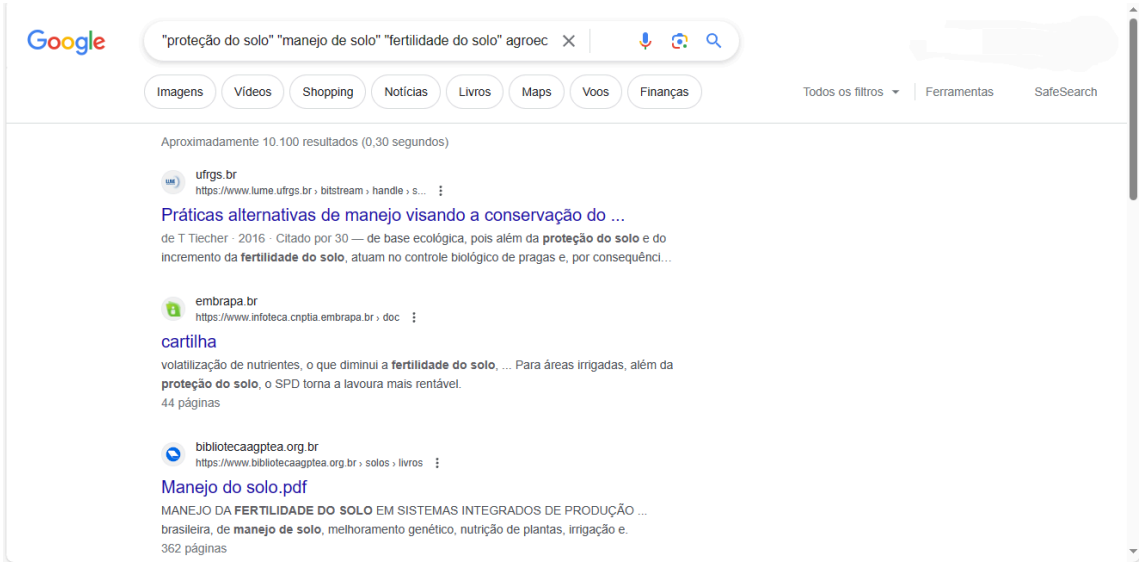
Figura 4: Resultado de busca da tupla *bioquímica do solo, manejo do solo, biomassa e agroecologia* nas páginas Web.



Fonte: Google

A terceira imagem das tuplas geradas pelo *BootCaT* exibe 10.100 de resultados de busca na *Web*:

Figura 5: Resultado de busca da tupla *proteção do solo, manejo de solo, fertilidade do solo e agroecologia* nas páginas Web.



Fonte: Google

Com o auxílio da pesquisa de *corpus,* os *corpora* desta pesquisa foram elaborados em dois estágios de extração*: corpus 1 e corpus 2* os quais extraíram grandes volumes de termos, porém no *corpus 1* as palavras e expressões não eram muito significativas para o cenário solo devido ao fato de apresentar termos muito técnicos. Com isso, leva-se em consideração apenas a extração dos termos e expressões do *corpus 2*, os quais foram encontrados nos textos da *web,* relevantes para o cenário Solo, e esses entrarão na macroestrutura do Dicionário, após análise quantitativa e qualitativa dos dados linguísticos do *corpora*.

Nas próxima seção, serão apresentados, como resultados da coleta dos dados do corpus 2, os termos e expressões mais relevantes e frequentes para entrar na Macroestrutura do Cenário SOLO os quais foram obtidos por meio da análise no software AntConc, assim como será mostrado um quadro da pesquisa com a definição do Cenário e as unidades lexicais.

6. RESULTADOS DA ANÁLISE DO *CORPUS* COMPILADO

As unidades lexicais selecionadas para compilar o primeiro *corpus* como *Microporos, Macroporos, Microagregados e Macroagregados* não foram relevantes para o domínio da Agroecologia, pois esses termos são muito técnicos. Visto que essas palavras estão mais presentes em textos acadêmicos, elas não entrarão para a Seleção Macroestrutural do Cenário SOLO, pois não estão de acordo com o tipo de dicionário que será construído e com os tipos de usuários dele.

Levando isso em consideração, foram escolhidas as palavras e expressões *qualidade do solo, proteção do solo, fertilidade do solo, biomassa e manejo do solo* que são representativas para a área da Agroecologia, pois essas unidades lexicais aparecem com bastante frequência no *corpus de apoio,* como os documentos da Embrapa, Emater e as Fichas Agroecológicas do MAPA. Tendo a análise qualitativa desses termos e expressões em vista, essas unidades lexicais foram inseridas como sementes no programa *BootCaT* para a construção do *corpus 2* .

Com isso, após a seleção das sementes e da combinação das sementes em tuplas do *corpus 2*, o programa forneceu textos extraídos da *web* em que elas aparecem. Além das sementes já fornecidas ao software, durante a construção do *corpus 2,* novas unidades lexicais foram coletadas nesses textos da internet. Esse segundo *corpus* foi aberto no *AntConc* e foram utilizadas as funcionalidades: *Word List*, *KWIC* e *Cluster*s para realizar a análise quantitativa das novas unidades. A

partir da inserção desse *corpus* no *AntConc*, primeiramente, foi verificada a lista de palavras e expressões, presentes no *corpus*, na guia *Word List* e, por meio dela, foram analisados os itens lexicais mais frequentes e significativos do *corpus,* em essas novas palavras e expressões aparecem, como: *solo, manejo, biomassa e sistema.*

Esses termos foram inseridos na funcionalidade *KWIC* para analisar como cada um desses itens lexicais aparecem nos textos ou nas frases, ou seja, para verificar o contexto de uso dessas unidades lexicais. Ao clicar no item lexical solo na funcionalidade *Word List*, o programa direcionou a busca para a *KWIC*, e essa exibiu como resultados palavras e expressões, além das que já foram fornecidas no *corpus* construído, como: *conservação do solo e qualidade do solo*. Já os termos *manejo, biomassa e sistema* apresentaram como resultados as expressões: *sistema de manejo de solo, biomassa seca e biomassa microbiana do solo.*

Na funcionalidade *Cluster* foi definida a palavra *solo* como termo de busca e foi ordenada para esquerda do *Cluster*. O *Cluster* apresentou como resultados dessa busca as expressões, além das que já foram inseridas no *corpus* construído.

A Tabela 1 abaixo apresenta a frequência dos termos e expressões coletados no *AntConc,* por meio das funcionalidades *Word List* e *Cluster:*

Tabela 1- Resultados dos termos e expressões na *Word List* e *Cluster*

Word List	Frequência na *Word List*	*Cluster*	Frequência no *Cluster*
Solo	9270.625	Qualidade do Solo(s)	1113
Manejo	1976.633	Fertilidade do Solo	705
Biomassa	544.035	Proteção do Solo	112
Sistema	1600.456	Cobertura do Solo	390
————	————	Preparo do Solo	227

Fonte: Elaborada pelo próprio autor.

Após a coleta dos termos mais frequentes e relevantes, foi elaborado um quadro com informações relevantes com a definição do Cenário SOLO, as unidades lexicais fornecidas ao *software Boot CaT* e as que foram coletadas no *Corpus 2:*

Quadro 1 - Resumo dos dados da coleta

CENÁRIO	**SOLO**	
Informações relevantes	**Definição**	**Complementares**
	O solo é um produto da matéria formado por sedimentos, rochas e matéria orgânica (restos de animais e plantas mortas) compostos de nutrientes, que são importantes para a manutenção dos seres vivos.	_____
Unidades lexicais utilizadas como sementes	agroecologia biomassa bioquímica do solo calagem corretiva fertilidade do solo manejo convencional manejo do solo proteção do solo qualidade do solo solo vivo	
Unidades lexicais retiradas do *corpus*	biomassa seca biomassa microbiana do solo cobertura do solo conservação do solo qualidade do solo preparo do solo sistema de manejo do solo	

Fonte: Elaborada pelo próprio autor.

7. CONSIDERAÇÕES FINAIS

Este trabalho apresentou a descrição das unidades lexicais do Cenário SOLO extraídas dos *corpora*, em que a seleção delas foi feita com base na macroestrutura de Bugueño (2006), e Atkins e Rundell (2008) as quais utilizam como parâmetro a frequência e a relevância dos itens lexicais para serem inseridos na macroestrutura, tendo em conta o tipo de usuário do dicionário e o dicionário que será criado. Seguindo esses métodos, foram compilados dois *corpora* no programa *BootCaT* e

um deles se mostrou representativo do Cenário SOLO, do qual foram extraídos diversos termos.

Após isso, levando em consideração o perfil de usuários, conhecimento enciclopédico e do tipo de dicionário que se almeja construir futuramente e do Domínio da Produção Rural, foram analisados e selecionados, no *AntConc*, apenas os itens lexicais relevantes e frequentes os quais são candidatos para entrar na macroestrutura do Dicionário. Isso é exemplificado a partir de alguns resultados apresentados na Tabela 1: *qualidade do solo, fertilidade do solo e biomassa seca*. Serão utilizados futuramente tanto para a elaboração dos Cenários e da macroestrutura do *Dicionário Enciclopédico da Agroecologia*, principalmente do Cenário SOLO, quanto para a descrição lexicográfica desses e das unidades lexicais os recursos lexicais da *FrameNet*. Essa aplicação de lexicografia computacional utiliza evidências baseadas em *corpus* e os preceitos da Semântica Cognitiva, a teoria da Semântica de *Frames,* o significado das palavras e expressões não é construído de forma isolada e é baseada na conhecimento enciclopédico, ou seja, para compreender e interpretar uma unidades lexical evocada por um *frame,* esta precisa estar relacionado ao Cenário SOLO e é necessário também ter o conhecimento enciclopédico vinculado às práticas de Produção Agroecológicas para o entendimento dela. Como exemplo disso: para que o usuário consiga compreender e interpretar a expressão *qualidade de solo*, esta *unidade lexical* deve estar conectada ao Cenário SOLO, e também é necessário o conhecimento enciclopédico (experiências e conhecimento de mundo armazenados na memória de longo prazo), voltado às práticas de Produção Agroecológica, do usuários para entendimento e compreensão do *frame*.

REFERÊNCIAS

ATKINS, B. T.; RUNDEL, Michael. **The Oxford guide to practical lexicography**. Oxford: OUP, 2008.

BIDERMAN, Maria Tereza Camargo. Introdução: as ciências do léxicos. *In*: OLIVEIRA, Ana Maria Pinto Pires de; ISQUERDO, Aparecida Negri(Orgs.). **As ciências do léxico**: lexicologia, lexicografia, terminologia. 2.ed. — Campo Grande, MS: Ed. UFMS, 2001.

CETESB.Companhia Ambiental do Estado de São Paulo. Definição de Solo. Disponível em: https://cetesb.sp.gov.br/solo/. Acesso em: 23 nov. 2023.

CHISHMAN, ROVE. **Dicionário Olímpico 2016**. 2016. Disponível em: https://dicionarioolimpico.com.br). Acesso em: 16 nov. 2023.

CHISHMAN, ROVE. **Dicionário Field.** 2018. Disponível em: https://www.dicionariofield.com.br. Acesso em: 16 nov. 2023.

EVANS, Vyvian; GREEN, Melanie. **Cognitive linguistics:** an introduction. Edimburgo: Edinburgh University Press, 2006.

FERRARI, Lilian. **Introdução à Linguística Cognitiva.** São Paulo: Editora Contexto, 2011.

FILLMORE, Charles. Frame Semantics. The Linguistic Society of Korea. Linguistic in the Morning Calm, Seoul, Hansinh Publishing Co, 1982.

FILLMORE, Charles. FrameNet Glossary. Disponível em: https://framenet.icsi.berkeley.edu/glossary. Acesso em: 16 nov. 2023.

KADER, Cárla Callegaro Corrêa; RICHTER, Marcos Gustavo. Linguística de corpus: possibilidades e avanços. **Instrumento: Revista de Estudos e Pesquisa em Educação***,* UFJF, v.15, n.3, 2013.

MARTELOTTA, Mário Eduardo; PALOMANES, Roza. Linguística Cognitiva. *In:* MARTELOTTA, Mário Eduardo (org.). **Manual de Linguística.** 2. ed — São Paulo: Contexto, 2021.

MIRANDA, Félix Bugueño. O que é macroestrutura no dicionário de língua? *In*: ISQUERDO, Aparecida Negri; ALVES, Ieda Maria. (Org.). **As ciências do léxico**: lexicologia, lexicografia, terminologia. V. III São Paulo: Humanitas, 2007.

MOREIRA FILHO, José Lopes. **Python para Linguística de *Corpus*:** Guia Prático. Editora Independente, 2021.

OLIVEIRA, Ana Flávia Souto de; ECHEVARRIA, Camile Heinrich. Estrutura de um dicionário enciclopédico do novo coronavírus organizado com base na semântica

cognitiva lexical: apontamentos sobre o domínio SAÚDE. **Revista do GEL (Grupo de Estudos Linguísticos do Estado de São Paulo)**, No Prelo.

OLIVEIRA, Ana Flávia Souto de. **SUBSÍDIOS DA SEMÂNTICA COGNITIVA PARA A DISPOSIÇÃO DAS ACEPÇÕES NOS LEARNER'S DICTIONARIES.** LUME/UFRGS. Disponível em: http://hdl.handle.net/10183/25440. Acesso em: 10 jun. 2023.

OLIVEIRA, Ana Flávia Souto ; REVELLES GATTI, Chrystian; PIPPER HATWIG, Guilherme. Utilização de corpora extraídos da web em um dicionário enciclopédico do novo coronavírus. **Revista Letras**. (62), 82–96. Disponível em: https://doi.org/10.5902/2176148568117. Acesso em: 10 jun. 2023.

SODRÉ, Fernando Fabriz. Química de Solos: Uma introdução". Artigos Temáticos do ACQUA(Grupo de Automoção, Quimiometria e Química Ambiental), n.1, 2012.Disponível em:https://www.aqqua.unb.br/images/Artigos/Tematicos/solos.pdf. Acesso em: 23 nov. 2023.

SALOMÃO, Maria Margarida Martins; TORRENT, Tiago Timponi; CAMPOS, Fernanda Cláudia Alves; BRAGA, Regina Maria Maciel. Dicionário FrameNet Brasil da Copa do Mundo. Disponível em: https://www.dicionariodacopa.com.br. Acesso em: 15 nov. 2023.

WELCKER, Helbert Andreas. **Dicionários:** Uma pequena introdução à lexicografia.Brasília:Thesaurus, 2004.