UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
CURSO DE BACHARELADO EM LINGUÍSTICA TEÓRICA E DESCRITIVA

Lucas Fonseca Lage

Mudanças Semânticas com Word Embeddings: um estudo de caso para o Português

Belo Horizonte
2021

Lucas Fonseca Lage

Mudanças Semânticas com Word Embeddings: um estudo de caso para o Português

Trabalho Conclusão do Curso de Graduação em Letras da Faculdade de Letras da Universidade Federal de Minas Gerais como requisito para a obtenção do Título de Bacharel em Linguística Teórica e Descritiva
Orientador: Prof. Dr. Evandro Landulfo Teixeira Paradela Cunha

Belo Horizonte

2021

Lucas Fonseca Lage

Mudanças Semânticas com Word Embeddings: um estudo de caso para o Português

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de Bacheral em Linguística Teórica e Descritiva e aprovado em sua forma final pelo Curso ...

Local, xx de xxxx de xxxx.

_____

Prof. xxx, Dr.

Coordenador do Curso

Banca Examinadora:

_____

Prof. Evandro Landulfo Teixeira Paradela Cunha, Dr.

Orientador

Universidade Federal de Minas Gerais

_____

Prof.ª Adriana Silvina Pagano, Dra.

Universidade Federal de Minas Gerais

_____

Prof. Bruno, Dr.

Universidade xxxxxx

Este trabalho é dedicado aos meus colegas de classe e aos meus queridos pais.

AGRADECIMENTOS

Inserir os agradecimentos aos colaboradores à execução do trabalho.

Texto da Epígrafe. Citação relativa ao tema do trabalho. É opcional. A epígrafe pode também aparecer na abertura de cada seção ou capítulo. Deve ser elaborada de acordo com a NBR 10520. (SOBRENOME do autor da epígrafe, ano)

RESUMO

De acordo com Givón (2001) o léxico é um repositório de conceitos relativamente estáveis no tempo, compartilhados socialmente e bem codificados, além de ser organizado em forma de rede, onde conceitos similares estão agrupados próximos uns aos outros. Em viés similar, o lexicólogo Georges Matoré propõe que palavras estabelecem relações associativas entre si e define os conceitos de campos nocionais e palavras-testemunho, elementos em torno dos quais o léxico se organiza. Com o uso de técnicas computacionais como *Word Embeddings*, que permitem a representação de palavras como vetores em um espaço vetorial, é possível analisar palavras agrupadas pelos mesmos traços semânticos. Esse trabalho se propõe analisar os contextos de ocorrência das palavras “Deus”, “Homem”, “Mulher”, “Pai”, “Mãe” e “Terra” no corpus Tycho Brahe do Português. Para isso utiliza-se o algoritmo *Skipgram* para gerar os *Word Embeddings*, e, posteriormente serem geradas visualizações para o campo semântico de cada palavra em três diferentes recorte temporais.

Palavras-chave: Processamento de Linguagem Natural. Diacronia. *Word Embeddings*. Mudança Semântica

ABSTRACT

Resumo traduzido para outros idiomas, neste caso, inglês. Segue o formato do resumo feito na língua vernácula. As palavras-chave traduzidas, versão em língua estrangeira, são colocadas abaixo do texto precedidas pela expressão “Keywords”, separadas por ponto.

Keywords: Keyword 1. Keyword 2. Keyword 3.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

NLP – Natural Language Processing

SUMÁRIO

Índice

1 INTRODUÇÃO

O presente trabalho tem como objetivo principal avaliar as mudanças semânticas sofridas por expressões ou palavras através de técnicas desenvolvidas pela área de Processamento de Linguagem Natural (NLP na sigla em inglês).

Tendo em vista os recentes desenvolvimentos de técnicas na área de computação, em especial as técnicas de Word Embeddings, novos tipos de análise linguística tem se tornado possíveis. Juntamente com a área de linguística de corpus, que tem exaustivamente criado dados de qualidade, tornam-se viáveis novas metodologias de pesquisa que permitem analisar uma grande quantidade de dados. (SARDINHA, 2004)

As técnicas de manipulação de dados muitas vezes não precisam ser sofisticadas, como vemos no trabalho de Michel et al. (2011), onde, através de medidas de frequência e análise de entidades culturalmente relevantes, são expostos fenômenos sobre a evolução da gramática, relevância cultural, censura e a construção de dicionários. (MICHEL et al., 2011)

A fundamentação teórica para a pesquisa seu pautará em uma abordagem lexicológica e funcional, tendo em vista que os conceitos aplicados de forma prática pelos algoritmos de *Word Embeddings* foram fundamentados teoricamente muito antes por autores como Georges Matoré e Talmy Givón. Buscando uma metodologia voltada para a sociolinguística, as transformações semânticas serão analisadas quanto as variáveis sociais relevantes, como idade, classe/status social, religião, contexto de uso e meio, sempre que possível. (GIVÓN, 1995) (CAMBRAIA, 2013)

Como fonte de dados será utilizado o corpus diacrônicos do português brasileiro Corpus Tycho Brahe, organizado pela Unicamp, que abrange textos do século XIII até o século XX. Além desse poderão ser usados outros corpora, desde que considerados adequados e que as técnicas permitirem. (DE SOUSA, 2014)

2 OBJETIVOS

O trabalho visa, de maneira geral, realizar uma análise sociolinguística de mudança semântica diacrônica. Entretanto, esses não seriam os únicos méritos do trabalho. Espera-se divulgar as metodologias mais recentes desenvolvidas em outras áreas para o campo da linguística e, também, contribuir com conhecimentos desenvolvidos no campo da linguística para a área de Processamento de Linguagem Natural.

São então os objetivos:

- Análise de mudanças semânticas com o auxílio de técnicas de NLP;
- Avaliar a viabilidade de técnicas recentes de computação em estudos linguísticos;
- Divulgação do uso de técnicas em computação para análise linguística;
- Intensificar a participação de profissionais de linguística no desenvolvimento de sistemas de processamento de linguagem natural

3 REVISÃO BIBLIOGRÁFICA

3.1 GIVÓN E O FUNCIONALISMO

Segundo Cunha, diferentemente das correntes estruturalistas e gerativistas, que buscam uma separação clara entre a língua como sistema (*langue*, *competência*) e a língua em uso (*parole*, *desempenho)*, o funcionalismo trata a estrutura gramatical da língua em relação ao seus contextos de uso. Dessa forma, as pesquisas dessa vertente se diferenciam de seus predecessores nos métodos utilizados, nos dados considerados relevantes e, mais profundamente, nos objetivos da análise linguística.(CUNHA, 2008)

Dentro de uma perspectiva funcionalista, uma sentença seria analisada sempre de acordo com o contexto onde ela foi produzida. Dessa forma, pode-se afirmar que a metodologia de análise parte de um método indutivo, analisando os dados, criando generalizações e, só então, testando essas generalização. Portanto, os dados analisados por uma pesquisa funcionalista devem ser dados e produções reais de fala e escrita. (CUNHA, 2008)

Dentro da perspectiva funcionalista destacam-se os estudos de Givón(1995, 2001), que desenvolveu uma gramática propriamente funcionalista. Givón, em sua obra, traz conceitos de outros campos, como a biologia, para justificar os caminhos tomados pela evolução das línguas. Segundo o autor, a evolução biológica é cercada por inúmeros fatores, muitas vezes envolvendo fenômenos aleatórios, e a forma que persevera é a forma que melhor realiza uma função específica. Essa evolução, considerada funcional, ocorre de forma similar nas línguas naturais. Givón ainda afirma que, apesar de os pontos abordados por ele não serem novos, como por exemplo, a capacidade humana de processamento de linguagem ser uma evolução do sistema de processamento de imagens visuais, a abrangência de fatos que são são influenciados por essas conclusões não foi ainda estudada. (GIVÓN, 1995, 2001)

Em seu livro *Funcionalism and Grammar*, o autor lista os componentes funcionais para a comunicação humana e os divide em dois módulos principais, que interagem entre si. São eles o *Sistema de Representação Cognitiva* e os *Sistemas de Codificação*. Dentro do Sistema de Representação Cognitiva temos três componentes, são eles: o léxico conceitual, a informação proposicional e o discurso multi proposicional. Aqui, nos interessa principalmente a definição de léxico conceitual, pois é a partir dela que Givón relaciona, inicialmente, o léxico ao meio e as experiências humanas.(GIVÓN, 1995)

O léxico humano é definido na obra *Functionalism and Grammar* como um conjunto de conhecimentos que, quando tomados juntos, constituem um mapa cognitivo do nosso universo de experiências como seres humanos. Esse universo de experiências se refere aos meios

externo-físico, ao universos socio-cultural, e ao nosso universo mental-interno. Além disso, os conceitos que compõe o léxico são definidos por Givón como estáveis no tempo, compartilhados socialmente e bem codificados. (GIVÓN, 1995)

Nessa visão, ser estável no tempo significa que as palavras e os conceitos associados a elas não estão em um fluxo rápido, ou seja, o termo "cavalo" provavelmente terá o mesmo referente daqui a alguns anos. Dizer que os conceitos são compartilhados socialmente significa que as palavras possuem aproximadamente o mesmo significado para os outros membros de sua comunidade de fala. E, por fim, ser bem codificado quer dizer que cada parte da informação armazenada no léxico é, em partes, fortemente associada a apenas um código, ou etiqueta perceptual. Ou seja, cada parte do conhecimento lexical possui apenas um correspondente no código.(GIVÓN, 2001)

Com essas características do léxico, Givón conclui que ele está organizado através de nós e conexões, e que cada nó corresponde a uma palavra. A ativação de um nó-palavra ainda seria responsável pela ativação outros nós-palavra que possuam uma relação íntima com o primeiro. Dentro de uma rede léxico-semântica, os nós correspondem a conceitos individuais, cada um com seu próprio significado e código-etiqueta. As conclusões de Givón são justificadas pelo trabalho de Swinney (1979), que analisa o tempo de reconhecimento de palavras, quando apresentadas a um leitor em contextos ambíguos, e conclui que, quando uma palavra é percebida, todos os possíveis significados dela são também ativados na mente da pessoa. (GIVÓN, 2001) (SWINNEY, 1979)

Os conceitos lexicais são as experiências humanas armazenadas de forma convencional e genérica, e não pontos específicos para cada experiência. Por serem genéricos, eles presumem um padrão de ativação para os conjuntos de nós interconectados. Um conceito lexical pode se referir a uma entidade relativamente estável no tempo, como um objeto, uma cidade, um local, animal ou até a conceitos abstratos, essa entidade corresponde então aos substantivos. Pode se referir a uma ação temporária, um processo ou relação, ou seja, um verbo. E por fim, pode representar uma qualidade estável no tempo ou temporária, como um adjetivo. (GIVÓN, 2001)

A ideia do léxico em rede descrita por Givón, como ele mesmo coloca, não é necessariamente nova. Outros trabalhos já tentaram trabalhar o léxico de forma sistemática em outros tempos a partir de outras perspectivas. Na próxima seção veremos como Georges Matoré deu início ao conceito do léxico em rede buscando uma lexicologia social.

3.2 LÉXICO EM REDE

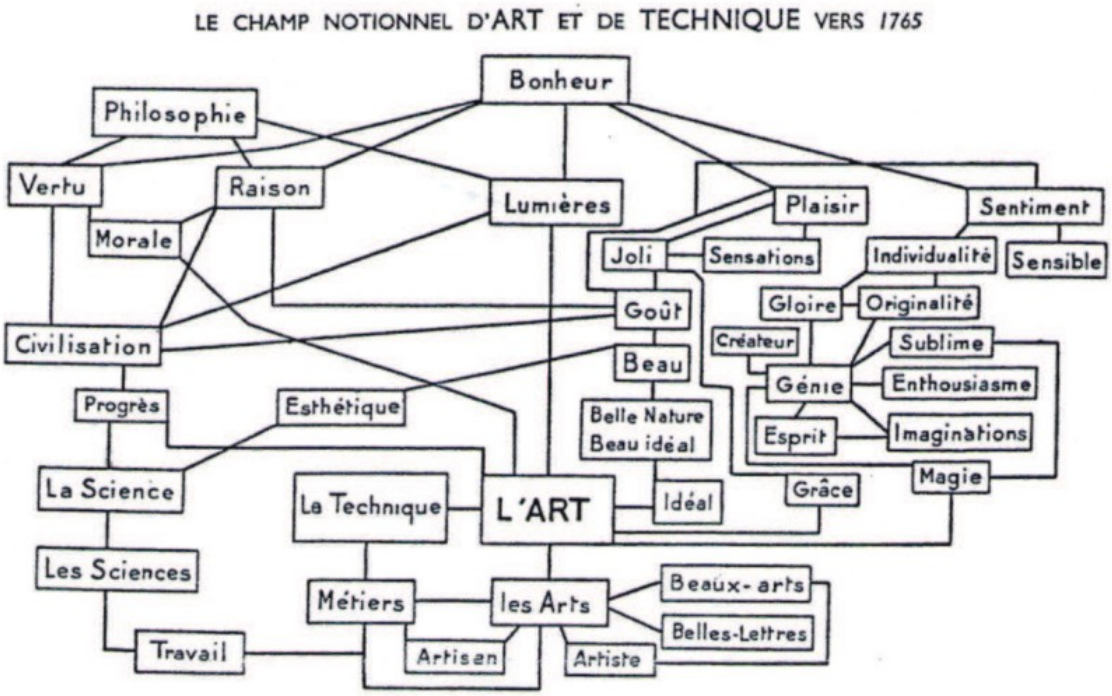
Através de seu artigo *La lexicologie sociale*, Georges Matoré se torna um dos primeiros a citar a influência de fatores sociais no estudo do léxico. Ele propõe uma série de princípios para a uma nova lexicologia, denominada lexicologia social. Dentre eles o primeiro princípio propõe que forma e conceito são indissociáveis. Segundo o autor, a formação de uma palavra equivale a formação de um conceito e esse processo criativo, apesar de individual em seu início, é seguido de uma socialização, que difunde e coletiviza a palavra e o conceito. Portanto, existe um caráter social da palavra e é por esse aspecto da significação que a lexicologia deveria se interessar.(CAMBRAIA, 2013)

A interpretação de Matoré sobre como o vocabulário se comporta é sistêmica, ou seja, admite que as palavras estabelecem relações recíprocas na consciência. As palavras podem se relacionar com suas vizinhas, através de relações sintagmáticas, ou com palavras similares, através de forma ou sentido, estabelecendo relações associativas. Matoré ainda afirma ser impossível extrair o fator tempo de suas análises, pois o momento de criação da palavra faz parte do conjunto de operações mentais que a produziram. (CAMBRAIA, 2013)

A metodologia de estudo do francês define que se façam recortes temporais que levem em conta a noção de "geração", cuja definição é uma faixa de tempo de 30 a 36 anos. Em seguida, devem ser identificados os *campos nocionais*, baseados no parentesco sociológico dos elementos. Esses campos são compostos por *palavras-testemunho*, que são elementos importantes em torno dos quais a estrutura lexicológica, sua hierarquia e sua coordenação são estabelecidos. Com base nesses métodos, Matoré exemplifica seu estudo através dos campos nocionais de Arte e Técnica por volta de 1765, figura 1, e o campo nocional de Artista entre os anos de 1827 e 1834, figura 2. (CAMBRAIA, 2013)

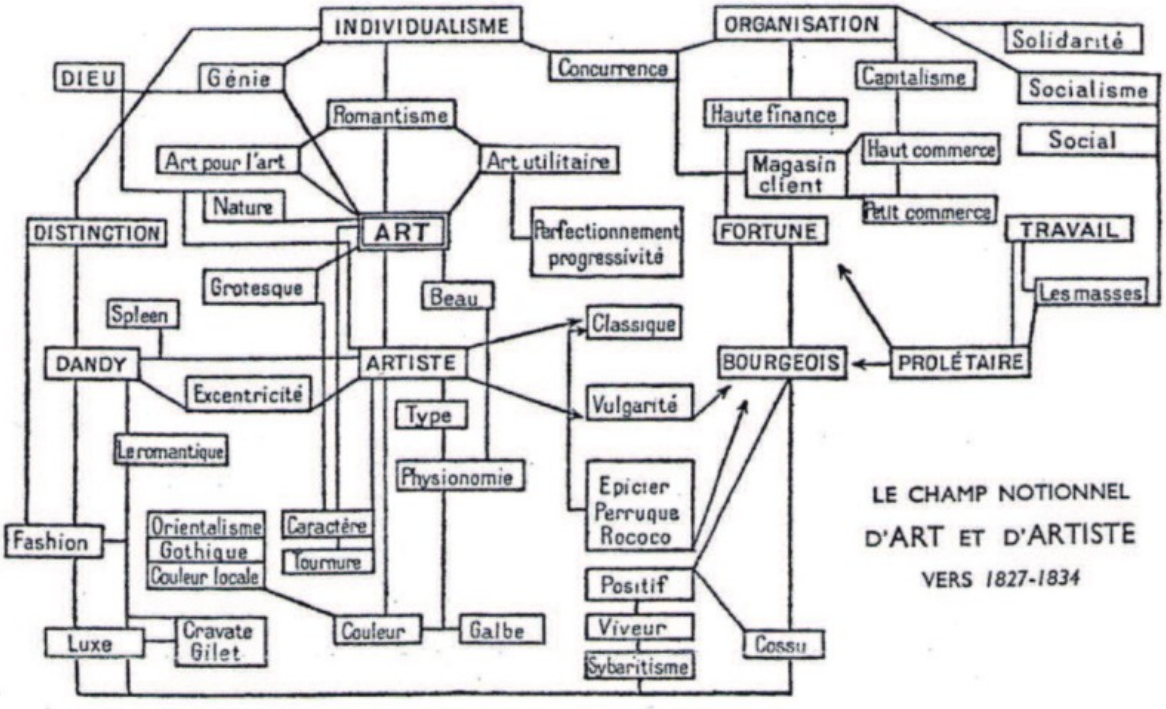
Os métodos defendidos por Matoré foram muito criticados para a época. Algumas críticas eram voltadas a partes mais técnicas do trabalho, como a definição arbitrária de uma geração de 30 a 36 anos, ou a imprecisão na definição de termos "palavras-testemunho" e "campos nocionais". Além disso, uma consideração muito importante foi realizada por Robin. A autora considera, aos criticar os métodos, que os estudos de Matoré não poderiam refletir a sociedade como um todo, mas apenas nos grupos a qual pertencem as pessoas cujos textos foram analisados. Apesar dos problemas da metodologia proposta pelo lexicólogo, permanece em destaque a importância de se considerar a influência do social na organização do léxico. (ROBIN; DE MENESES BOLLE, 1977) (CAMBRAIA, 2013)

Figura 1: Campo nocional de "Arte" e "Técnica" em 1765



Fonte:(CAMBRAIA, 2013)

Figura 2: Campo nocional de "Artista" entre 1827-1834.



Fonte: (CAMBRAIA, 2013)

Entretanto, os estudos de Matoré não podem ser considerados de cunho sociolinguístico, pois não levam em consideração elementos como classe social, idade, gênero, formação escolar, localidade, etc. Essas limitações no desenvolvimento da Lexicologia Social serão levadas em contas nesse trabalho. (CAMBRAIA, 2013)

Com a definição do caráter funcional das palavras, as fortes relações desenvolvidas entre elas e o léxico ativado em rede, parece compreensível que a única limitação para um trabalho relevante nessa perspectiva seria uma metodologia que pudesse capturar essas complexidades.

3.3 WORD EMBEDDINGS

3.3.1 Semântica Vetorial

O conceito de *Word Embeddings*, assim como as propostas de Givón, também se desenvolveram a partir de conceitos evolucionários. Da mesma forma que espécies diferentes desenvolvem estruturas corporais similares por evoluírem em ambientes similares, palavras que ocorrem em contextos similares devem possuir significados similares. Essa hipótese, denominada hipótese distribucional, foi proposta por linguistas na década de 1950 quando eles perceberam que palavras sinônimas ocorrem no mesmo ambiente. (JURAFSKY; MARTIN, 2008)

Enquanto palavras não possuem muitos sinônimos, a grande maioria as palavras possui outras que são muito similares. Por exemplo, apesar das palavras "cão" e "gato" não serem sinônimas, há muitas semelhanças entre elas e os contextos onde elas ocorrem, ambas são substantivos relacionados a animais domésticos, por exemplo. A noção de similaridade é bastante útil para definir a similaridade entre duas sentenças, que é um componente importante de tarefas como resposta a perguntas, paráfrase e sumarização. (JURAFSKY; MARTIN, 2008)

Além de relações entre palavras como sinonímia, polissemia, antonímia e similaridade, palavras também possuem um caráter afetivo. O caráter afetivo, ou conotação, refere-se aos aspectos do sentido de uma palavra que estão relacionados aos sentimentos do falante ou do ouvinte. Assim, as palavras podem ter uma conotação positiva (feliz, bom, amor) ou uma conotação negativa (triste, mal, ódio). Um dos primeiros trabalhos sobre o sentido afetivo de palavras foi o de Osgood, onde ele cria três eixos para avaliar o sentido afetivo de uma palavra e então associa um valor numérico a cada eixo. Na figura 3 vemos as valorações em três eixos, valência, excitação e dominância. (JURAFSKY; MARTIN, 2008)

Figura 3: Variação quantitativa do sentido afetivo em três eixos

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

Fonte:(JURAFSKY; MARTIN, 2008)

A grande contribuição de Osgood para o campo foi a percepção de que as palavras poderiam ser representadas em um espaço vetorial a partir dos valores para cada eixo, criando assim, um espaço tridimensional para localização espacial das palavras.

3.3.2 Continous Bag of Words e Skip-Gram

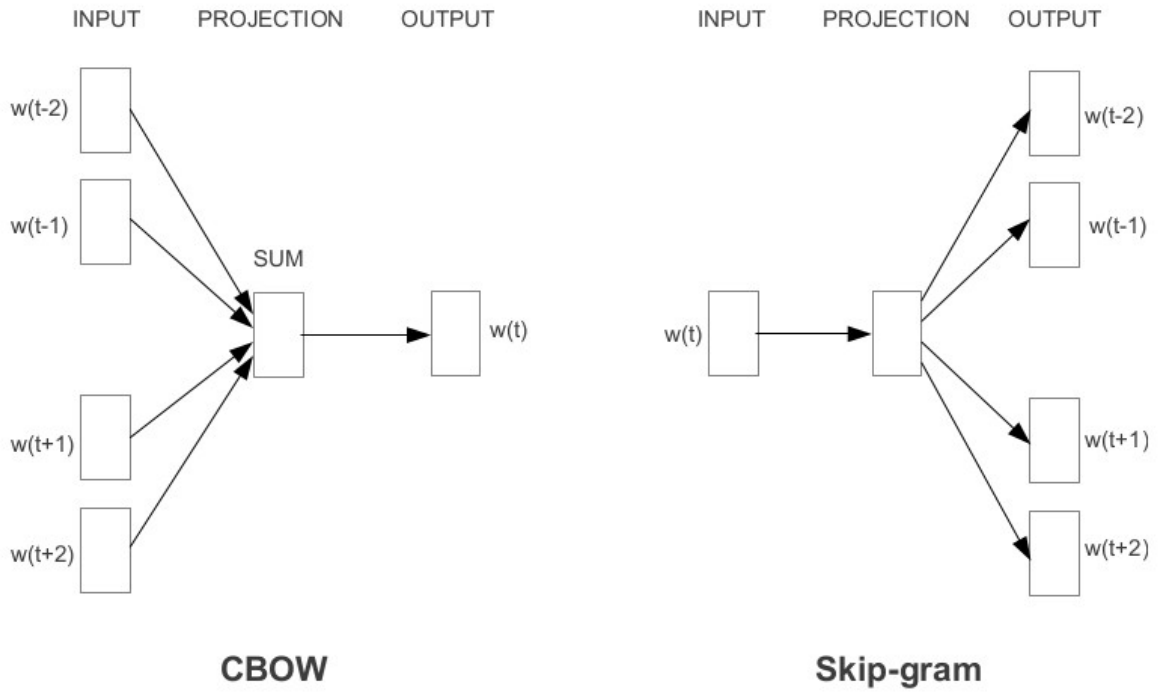
Um dos principais trabalhos com *Word Embeddings* é a publicação de Mikolov et. al. Nele é proposta uma arquitetura de redes neurais capaz de criar uma representação vetorial para cada palavra apresentada ao modelo, e, em seguida, reproduzi-las em um espaço vetorial. O mérito desse trabalho está relacionado a resolução de questões sobre a complexidade da arquitetura e o tempo necessário para o treinamento dessas redes. Além disso, também foram desenvolvidas métricas de validação de modelos que não só são capazes de determinar se palavras estão próximas entre si,como também de quantificar o grau de similaridade entre as palavras. (MIKOLOV et al., 2013)

As arquiteturas propostas pelos autores foram denominadas de *Continous Bag-of-Words (CBOW)* e *Continuous Skip-gram (Skip-gram)*. O primeiro deles é criado a partir de uma tarefa de predição, onde uma palavra é prevista dado seu contexto, ou palavras vizinhas, como entrada do modelo. O contexto, no caso, deve ser entendido como uma quantidade de palavras antes e depois da palavra a ser predita. A partir dos valores de entrada, um classificador *log-linear* calcula a palavra mais provável de ocorrer naquele contexto, caso a predição seja correta, a rede realiza operações dentro de si para reforçar seu aprendizado. Caso a predição esteja errada, ela altera valores dentro de si para buscar acertar nas próximas tentativas. É

importante ressaltar que a ordem das palavras dentro da janela de entrada não é um fator relevante para a predição. (MIKOLOV et al., 2013)

O modelo Skip-gram possui uma arquitetura similar ao modelo CBOW, mas ao invés de predizer uma palavra dado seu contexto, ele realiza a tarefa inversa, prediz o contexto a partir de uma palavra. (MIKOLOV et al., 2013)

Figura 4: Esquema da arquitetura dos modelos CBOW e Skip-gram



Fonte:(MIKOLOV et al., 2013)

Para validação dos modelos criados, os autores desenvolveram tarefas de predição baseadas em relações sintáticas e semânticas. Como exemplo de similaridade sintática, são utilizadas as relações entre os adjetivos do inglês em sua forma base, comparativa e superlativa. Essas relações podem ser preditas de acordo com os valores obtidos para os vetores após o treinamento do modelo. Assim, pode-se encontrar que a relação entre *big* e *bigger* é a a mesma que entre *small* e *smaller*. Essa relação pode então ser reescrita através de uma operação algébrica como Vetor("*big*") + Vetor("*bigger*") - Vetor("*small*") = Vetor("*smaller*"). (MIKOLOV et al., 2013)

Como exemplo de relação semântica foi criado um teste para determinar as relações entre nomes de países e suas capitais. Assim, podemos encontrar relações como "Paris está

para França assim como Berlim está Alemanha". Como um vetor essa operação seria como: Vetor("Paris") + Vetor("França") - Vetor("Alemanha") = vetor("Berlim"). (MIKOLOV et al., 2013)

Dentro dessas tarefas, o modelo Skip-gram atingiu a melhor acurácia total quando comparado com o modelo CBOW e outras arquiteturas comparadas. O desempenho geral do modelo CBOW foi a princípio ruim, mas teve o terceiro melhor desempenho dentro das tarefas de relações sintáticas. A grande vitória de ambos os modelos entretanto é no tempo de treinamento. Para as mesmas condições de treino, mesmo tamanho de corpus e mesmas capacidades de processamento, os modelos CBOW e Skip-gram demoraram 2 e 2,5 dias respectivamente para finalizaram o treinamento, enquanto as outras arquiteturas precisaram de 14 dias de treino para criar o seu modelo. (MIKOLOV et al., 2013)

Tendo em mãos uma técnica não só capaz de capturar relações sintáticas e semânticas, como também de fácil processamento, falta então a definição de um corpus relevante para se trabalhar.

3.4 DESCRIÇÃO DO CORPUS TYCHO BRAHE

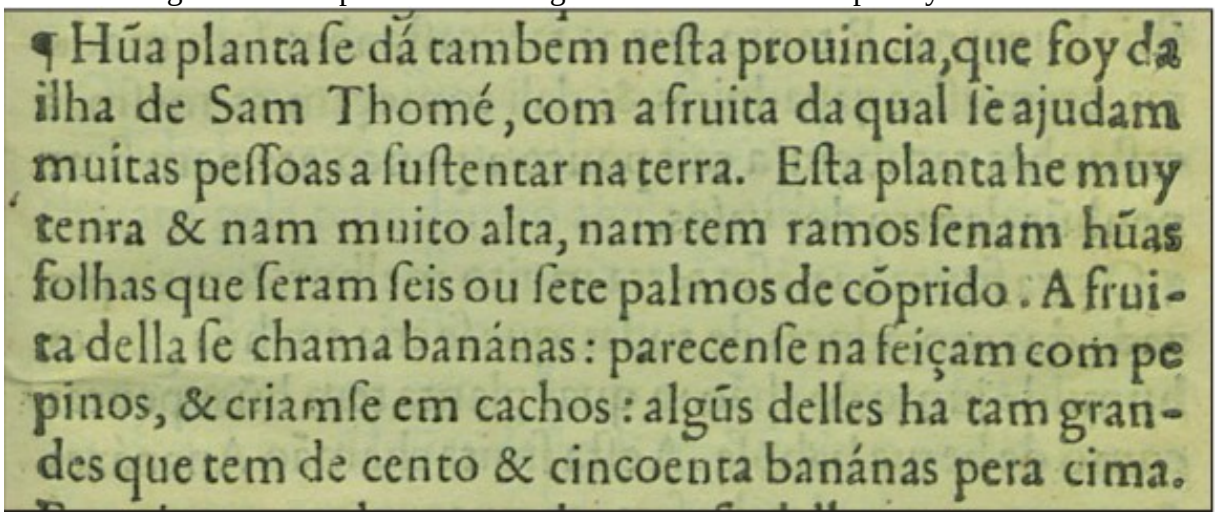
O Corpus Anotado do Português Histórico Tycho Brahe (CBT), construído a partir do projeto Galves (1998), foi pioneiro no que concerne a língua portuguesa, e permanece hoje como o maior corpus eletrônico anotado de textos históricos em português. Hoje, o conjunto de dados inclui textos escritos por autores portugueses, brasileiros e africanos, nascidos entre 1380 e 1845, publicados entre os anos 1350 e 1948. As anotações realizadas nos textos têm como objetivo principal possibilitar, de forma ampla, a recuperação de informações filológicas e linguísticas dos textos. (DE SOUSA, 2014)

O corpus é composto de 88 textos, totalizando 3.544.628 palavras, 58 textos anotados morfologicamente e 27 textos anotados sintaticamente. O corpus possui uma variedade de gêneros textuais, sendo eles: cartas, atas, textos narrativos, textos dissertativos, gramáticas, gazetas e jornais e textos de dramaturgia. (DE SOUSA, 2014)

Como muitos textos provém de séculos passados o processamento deles deve envolver uma adaptação para que possam ser lidos e analisados hoje. O processamento do texto a partir da obra original se dá então por 3 camadas, uma camada de edição, uma camada morfossintática e uma camada sintática. As anotações acontecem de forma incremental, ou seja, cada uma depende do resultado da etapa anterior.(DE SOUSA, 2014)

A primeira etapa é a anotação de edição, que codifica dois tipos de informação diferentes do texto original. A primeira delas são as informações relativas às decisões editoriais e à estrutura do texto (quebras de linha, parágrafos, seções, etc). A segunda lida com intervenções interpretativas, como atualização grafemática, expensão de abreviaturas, atualização ortográfica. A transcrição é feita com o auxílio de uma ferramenta chamada *e-Dictor*, que permite a realização de adequações de grafia. Por exemplo, no termo original *parecenſe*, é normalizado grafematicamente para *parecense*, e em seguida tem a grafia atualizada para *parecem-se*. Na figura 5 abaixo temos um exemplo de como a os textos originais se encontram. (DE SOUSA, 2014)

Figura 5: Exemplo de trecho original encontrado no corpus Tycho Brahe.


¶ Hũa planta ſe dá tambem neſta prouincia, que foy da
ilha de Sam Thomé, com a fruita da qual ſe ajudam
muitas peſſoas a ſuſtentar na terra. Eſta planta he muy
tenra & nam muito alta, nam tem ramos ſenam hũas
folhas que ſeram ſeis ou ſete palmos de cõprido. A frui-
ta della ſe chama banánas: parecenſe na feiçam com pe
pinos, & criamſe em cachos: algũs delles ha tam gran-
des que tem de cento & cincoenta banánas pera cima.

Fonte:(DE SOUSA, 2014)

A próxima etapa de anotação é a anotação morfossintática que consiste na identificação e codificação das classes de palavras. Para isso, foi usada a ferramenta *e-Dictor*, que conta com um classificador morfossintático com taxa de acerto de aproximadamente 95\%, realiza a anotação dos textos. No CTB, são usadas 381 etiquetas, que remetem a classes de palavras básicas (Nome, Verbo, Preposição, etc.), flexões (tempo, número) e aglutinações (Preposição + Determinante, etc.).

Figura 6: Exemplo de anotação morfossintática no corpus Tycho Brahe.

```
A/D-F fruta/N dela/P+PRO se/CL chama/VB-P bananas/N-P :/. parecem-se/VB-P+CL
na/P+D-F feição/N com/P pepinos/N-P,/, e/CONJ criam-se/VB-P+CL em/P cachos/N-
-P :/. alguns/Q-P deles/P+PRO há/HV-P tão/ADV-R grandes/ADJ-G-P que/C tem/
TR-P de/P cento/NUM e/CONJ cinqüenta/NUM bananas/N-P para/P cima/N ./.
```

Fonte:(DE SOUSA, 2014)

A terceira e última etapa é a anotação sintática. Nessa etapa é realizada a identificação e codificação da estrutura sintagmática da sentença. Para realizar a anotação sintática do corpus foi desenvolvido um parser sintático a partir do sistema Penn-Treebank. (DE SOUSA, 2014)

Figura 7: Exemplo de anotação sintática no corpus Tycho Brahe.

```
( (IP-MAT (NP-SBJ *exp*)
          (NP-ACC (Q-P alguns)
                  (PP (P d@)
                      (NP (PRO @eles))))
          (HV-P há)
          (ADJP (ADV tão)
                (ADJ-G-P grandes)
                (CP-DEG (C que)
                        (IP-SUB (NP-SBJ *pro*)
                                (TR-P tem)
                                (PP (PP (P de)
                                        (NP (NUMP (NUM cento)
                                                  (CONJ e)
                                                  (NUM cinqüenta))
                                            (N-P bananas)))
                                    (P para)
                                    (NP (N cima))))))
          (. .))  (ID G_008,17.204))
```

Fonte:(DE SOUSA, 2014)

O parser sintático recebe como entrada as análises prévias do classificador morfossintático e um texto com sua estrutura sintática manualmente anotada. O parser em seguida realiza as operações necessárias para a predição e as otimizações baseadas em seus erros e acertos. (DE SOUSA, 2014)

Por fim, a aplicação de técnicas de *Word Embeddings* em corpora diacrônico veio um pouco mais tarde com o trabalho realizado por Hamilton, Leskovec e Jurafsky (2016). (HAMILTON; LESKOVEC; JURAFSKY, 2016)

3.5 WORD EMBEDDINGS E CORPORA DIACRÔNICO

Hamilton, Leskovec e Jurafsky (2016), propõe em seu trabalho duas leis para a mudança semântica. São elas a Lei da Conformidade e a Lei da Inovação. A primeira diz que a velocidade com que uma palavra muda seu sentido é inversamente proporcional a uma função

exponencial da frequência de palavras. Já a segunda, alega que, dentre palavras com frequência de ocorrência similar, as palavras polissêmicas mudam seu sentido mais rapidamente.(HAMILTON; LESKOVEC; JURAFSKY, 2016)

Para chegar a essa conclusão os autores utilizam três diferentes arquiteturas de *Word Embeddings* e corpora diacrônicos que englobam quatro línguas diferentes, sendo elas inglês, alemão, francês e chinês. Foram então criados modelos de *Word Embeddings* que abrangiam diferentes períodos de tempo e, após alinharem os modelos para cada período, foi criada uma representação para palavras cuja mudança semântica é conhecida. As palavras escolhidas foram *broadcast*, *gay* e *awful* e a partir das mudanças sofridas obteve-se uma representação gráfica visível na figura 1. (HAMILTON; LESKOVEC; JURAFSKY, 2016)

Figura 8: Deslocamento num espaço vetorial das palavras broadcast, gay e awful entre1800 e 1999.

Fonte:(HAMILTON; LESKOVEC; JURAFSKY, 2016)

A escolha das palavras pelos autores não foi aleatória, foram escolhidas palavras que pudessem validar a metodologia descrita por eles. Assim, esperava-se que a palavra *broadcast* estivesse ligada a termos relacionados a agricultura em um primeiro momento, e, em seguida, estivesse próxima de termos relacionados a notícias, jornais, televisão e radio. (HAMILTON; LESKOVEC; JURAFSKY, 2016)

O estudo conseguiu, em um primeiro momento, validar mudanças de significado já conhecidas e foi em seguida usado para buscar as palavras que sofreram maior mudança semântica, aqui considerado o maior deslocamento no espaço vetorial ao longo dos períodos analisados.

4 METODOLOGIA

A partir da visão de Givón sobre o léxico e sobre a lexicologia social proposta por Matoré, propomos no presente trabalho buscar a representação da mudança semântica de palavras relevantes para o português. Para isso será utilizado o Corpus Tycho Brahe, elaborado por Sousa (2014), que ainda é referência nos estudos diacrônicos para o português.

A princípio foi realizada uma análise exploratória. Nela foram encontradas as palavras de maior frequência e qual é a distribuição das palavras do corpus.

Buscando uma metodologia similar ao de Hamilton et, o corpus será divido em períodos relevantes para a mudança semântica. Entretanto, a determinação de um período relevante se mostra bastante imprecisa. Além disso, apesar da importância do Corpus Tycho Brahe, a quantidade de palavras no corpus é muito menor comparada aos corpora utilizados por Hamilton et al(2016). A fim de minimizar o impacto desses obstáculos, os textos foram agrupados por século, garantindo uma quantidade minimamente significativa de tokens para cada século e mantendo um recorte de tempo padronizado. Além disso, há uma escassez de textos nos períodos mais antigos da língua, em especial no século XIV. Por esse motivo serão considerados apenas textos a partir do século XV. (HAMILTON; LESKOVEC; JURAFSKY, 2016)

A análise inicial foi feita em palavras que foram consideradas relevantes cultural e socialmente pelo autor, consideradas através da perspectiva de palavras-testemunho e campos nocionais de Matoré. Dessa forma, serão buscados a mudança de termos como *homem, mulher, pai, mãe e Deus*. Espera-se encontrar palavras que possam dar pistas sobre a percepção e os conceitos culturais que permearam esses termos durante a história da língua portuguesa.

A ferramenta principal utilizada para essa análise foi a linguagem de programação Python, além das diversas bibliotecas para NLP existentes como Spacy e NLTK (Natural Language Toolkit).

4.1 ANÁLISE EXPLORATÓRIA

Para a análise exploratória o corpus foram feitas as seguintes etapas de pré-processamento:

- Tokenização;
- Remoção de Stopwords;
- Remoção de acentuação

Após realizados esses passos foi obtido o número de ocorrências das palavras mais frequentes, podendo ser visto na tabela 1:

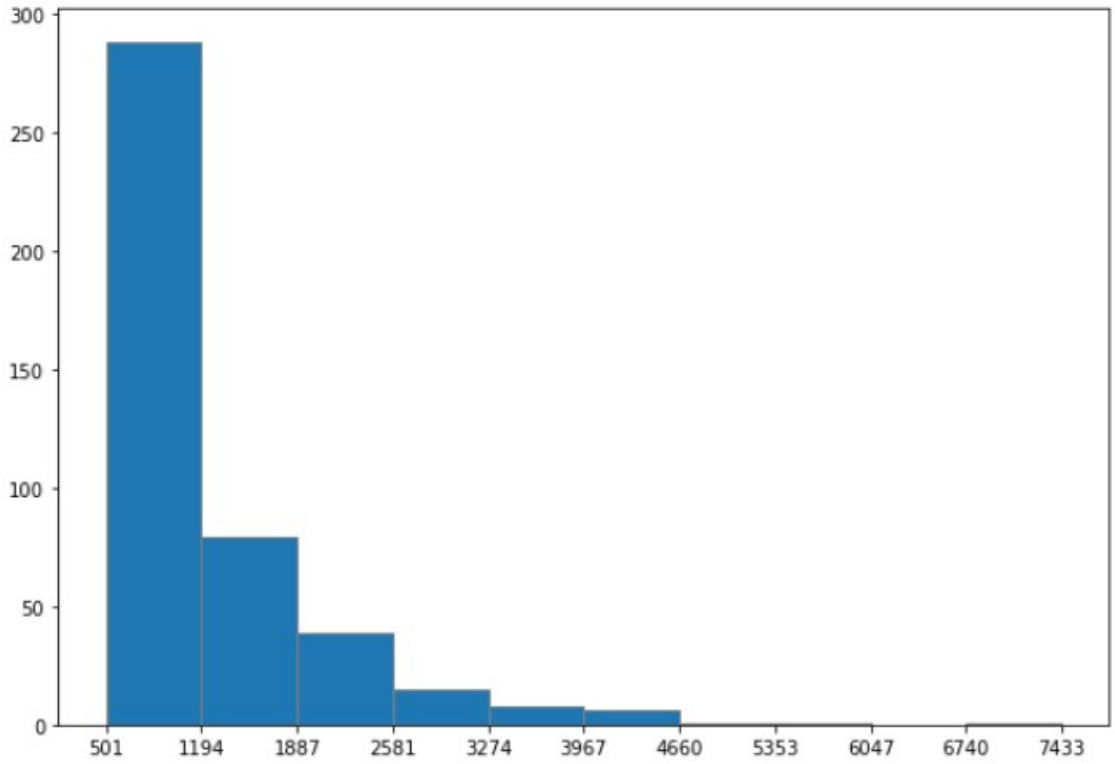
Tabela 1: Palavras mais frequentes do corpus

Palavra	Número de Ocorrências
senhor	7433
bem	6005
deus	4602
grande	4600
dom	4589
assim	4584
tempo	4265
tudo	4126
pois	3922
fazer	3738

Fonte: Elaborado pelo autor

A distribuição das frequências do corpus pode ser vista na figura 9 a seguir:

Figura 9: Distribuição das frequências do corpus



Fonte: Elaborado pelo autor

Por fim foram encontradas as frequências para as palavras selecionadas para a análise.

Tabela 2: Número de ocorrências das palavras a serem analisadas

Palavra	Frequência
pai	1312
mãe	798
deus	4602
homem	2506
mulher	1229
terra	2209

Fonte: Elaborado pelo autor.

4.2 DEFINIÇÃO DOS PERÍODOS

A separação do corpus em períodos se dá pela necessidade de sistematizar o agrupamento dos textos em relação ao período em que ele ocorre. O agrupamento foi realizado de acordo com a classificação proposta por Bechara (1985). (BECHARA, 1985)

De acordo com o autor as fases do português podem ser definidas de acordo com os seguintes agrupamentos:

Tabela 3: Classificação de períodos do Português e nr. de Tokens

Fases	Séculos	Nr. de Tokens
Arcaica / Arcaica Média	Até 1ª metade séc. XVI	632.907
Moderna	2ª metade séc. XVI até fim séc. XVII	1.230.507
Contemporânea I	Séc XVII	615.108
Contemporânea II	Séc XIX e início séc. XX	824.289

Fonte: Elaborado pelo autor.

Como o número de tokens para as fases Contemporânea I e II são menores, foi decidido juntar os dois grupos em um para que o número de tokens entre a fase moderna e contemporânea fosse balanceado. Dessa forma temos o seguinte agrupamento final para ser analisado, mostrado na tabela.

Tabela 4: Agrupamento final dos textos de acordo com seu período.

Agrupamento	Séculos	Nr. de Tokens
Período I	Até 1ª metade séc. XVI	632.907
Período II	2ª metade séc. XVI até fim séc. XVII	1.230.507
Período III	Séc XVII até início séc. XX	1.439.397

Fonte: Elaborado pelo autor.

Os textos foram em seguida agrupados de acordo com a data de nascimento do autor e então tratados para o posterior treinamento do modelo.

4.2 LIMPEZA E PROCESSAMENTO DO CORPUS

Após o agrupamento dos textos, foram realizados os seguintes passos de pré processamento:

- *Tokenização*;
- Transformação do texto em caixa baixa;
- Retirada de espaços em branco em excesso;
- Retirada de *Stopwords*;
- Lematização;
- Retirada de acentos gráficos;

A *tokenização* consiste quebrar o texto em pedaços menores que possuam relevância para a análise, esses pedaços podem ser palavras, símbolos gráficos ou numerais, desde que sejam relevantes para a análise. O resultado da *tokenização* é o *token* ou uma lista de *tokens.*

Em seguida o texto é transformado completamente em caixa baixa, evitando assim que *tokens* iguais sejam considerados diferentes devido a escrita em maiúscula. Dessa forma, os *tokens* “Portanto” e “portanto” são transformados no mesmo *token* “portanto”.

A retirada de espaços em branco em excesso se dá para padronização dos espaçamentos e facilitar o processamento dos arquivos de texto.

A retirada de *Stopwords* é um processo comum em tarefas de NLP. As *stopwords* são palavras consideradas gramaticais ou pouco relevantes semanticamente. Sendo assim elas são retiradas em análises de NLP. Existem diversas listas de *Stopwords* disponíveis com diferentes critérios, neste trabalho foi usada a lista fornecida por padrão pela biblioteca NLTK.

A lematização é um processo também comum em tarefas de NLP e consistem deflexionar uma palavra, retornando ela para sua forma base, dicionarizada. O resultado desse processo é que substantivos estarão no singular e sem flexão de gênero, verbos estarão no infinitivo e adjetivos estarão sem flexão de gênero.

Por último a retirada de acentos gráficos é feita para evitar que o *encoding* do texto afete o resultado.

Todos os passos foram feitos através da biblioteca Spacy.

4.3 TREINAMENTO DOS MODELOS

Os modelos Skipgram foram treinados, um para cada agrupamento, através do código fornecido por (“tmikolov/word2vec”, [s.d.]).

Os hiperparâmetros são as condições de treinamento utilizadas, e forammant mantidas no padrão. O motivo dessa escolha é justificado pela diminuição nos ganhos com a alteração dos parâmetros, como mencionado por Mikolov et al. (MIKOLOV et al., 2013)

Os hiperparâmetros mais relevantes de treinamento podem ser vistos na tabela seguinte.

Tabela 5: Hiperparâmetros de treinamento

Hiperparâmetro	**Valor**
Tamanho do vetor (size)	300
Janela (window)	8
Amostragem negativa (negative)	25
Amostra (sample)	1e-4
Binário (binary)	1
Iterações (iter)	25

Fonte: Elaborado pelo autor.

O parâmetro *size* diz respeito ao tamanho do vetor, ou número de dimensões que o vetor de cada palavras possuirá após o treinamento. O parâmetro *window* se refere a janela de treinamento, o seu valor determina o número de *tokens* antes e número de *tokens* depois da palavra alvo, para um intervalo total de 16+1. O parâmetro *negative* corresponde a amostragem negativa, que é o número de exemplos negativos gerados para o treinamento. O valor fornecido corresponde a razão entre o número de amostras negativas e positivas, no nosso caso temos 25 vezes mais amostras negativas que positivas. Por fim o parâmetro *iter* diz respeito ao número de iterações necessárias para o treinamento, no caso o treinamento foi repetido 25 vezes.

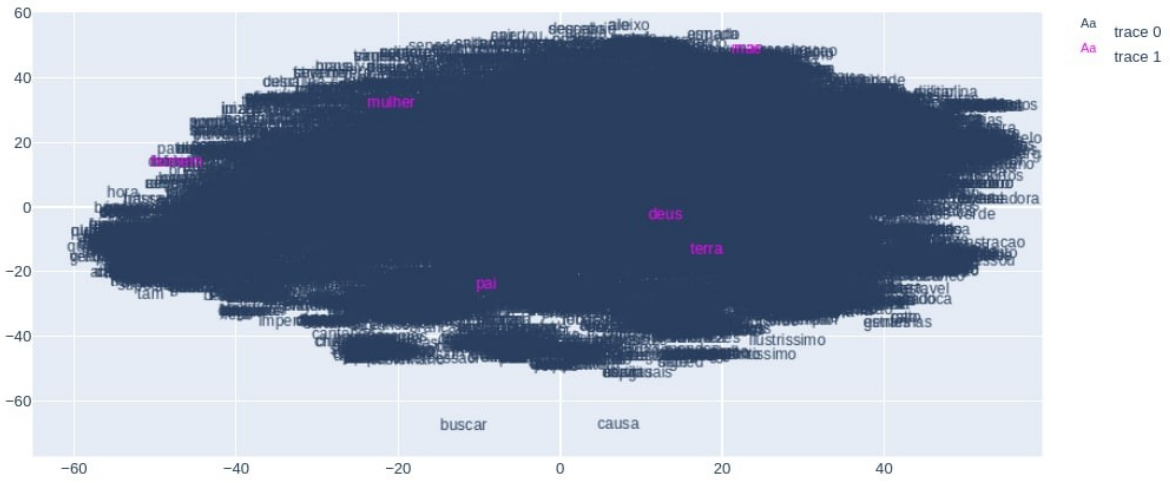
Após a obtenção dos vetores de palavras resta apenas gerar a visualização gráfica.

4.4 GERAÇÃO DE VISUALIZAÇÕES

Por fim foram geradas visualizações para a representação dos vetores n-dimensionais em um espaço bidimensional. Para isso foi utilizado a biblioteca Python, Gensim.

Com ela, foi usado do algoritmo T-SNE para a redução dos vetores de 300 dimensões a apenas duas dimensões e por fim, os pontos foram plotados em um espaço bidimensional, como mostra a figura 10 a seguir.

Figura 10: Representação gráfica gerada pelo Gensim.



Fonte: Elaborado pelo autor.

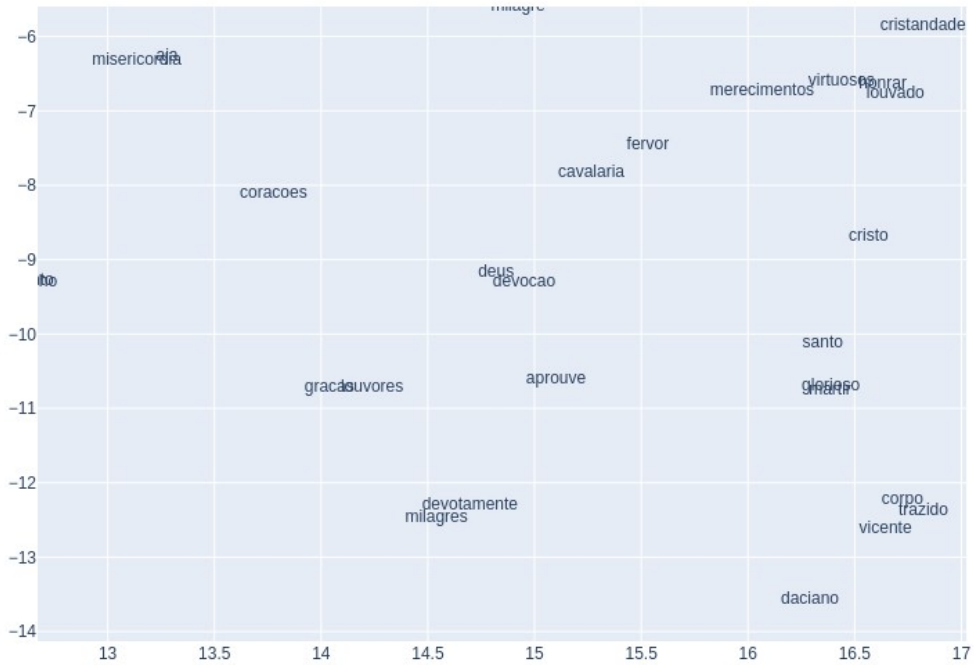
5 RESULTADOS

5.1 CAMPOS SEMÂNTICOS PARA PALAVRA “DEUS”

Como primeiro resultado obteve-se os campos semânticos das palavras escolhidas inicialmente para cada um dos períodos delimitados no corpus.

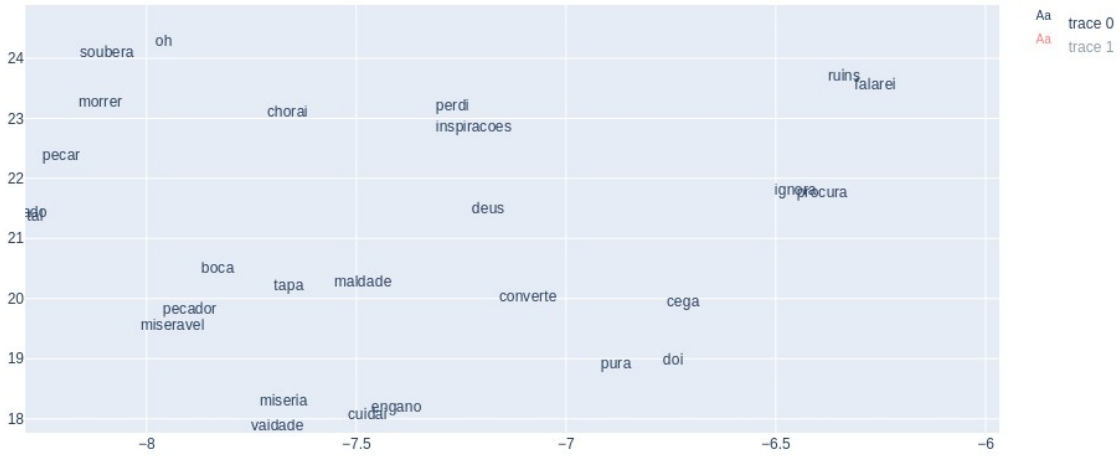
Primeiramente, os gráficos gerados para a palavra “Deus” no período I mostram-se adequados ao contexto, com palavras como “devoção”, “louvores”, “devotamente” (figura 11), em sua proximidade. Já para o período II, vemos palavras como “inspirações”, “converte”, “perdi” e “maldade” (figura 12) próximas. Por fim, no período III têm-se as palavras “amas”, “aflitos”, “misericórdia”, “divinos”, “preconceito”, “santíssima”. (figura 13).

Figura 11: Campo semântico da palavra "Deus", período I



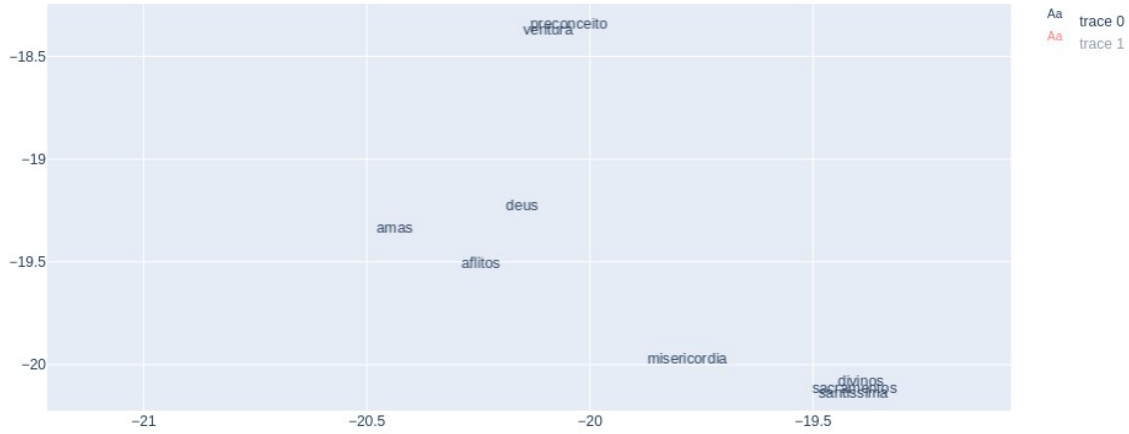
Fonte: Elaborado pelo autor.

Figura 12: Campo semântico da palavra Deus, período II



Fonte: Elaborado pelo autor.

Figura 13: Campo semântico da palavra "Deus", período III

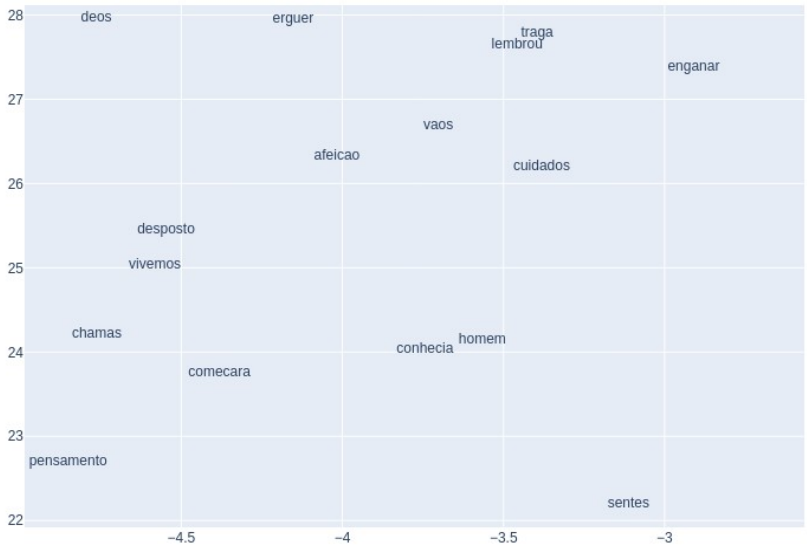


Fonte: Elaborado pelo autor.

5.2 CAMPOS SEMÂNTICOS PARA PALAVRA “HOMEM”

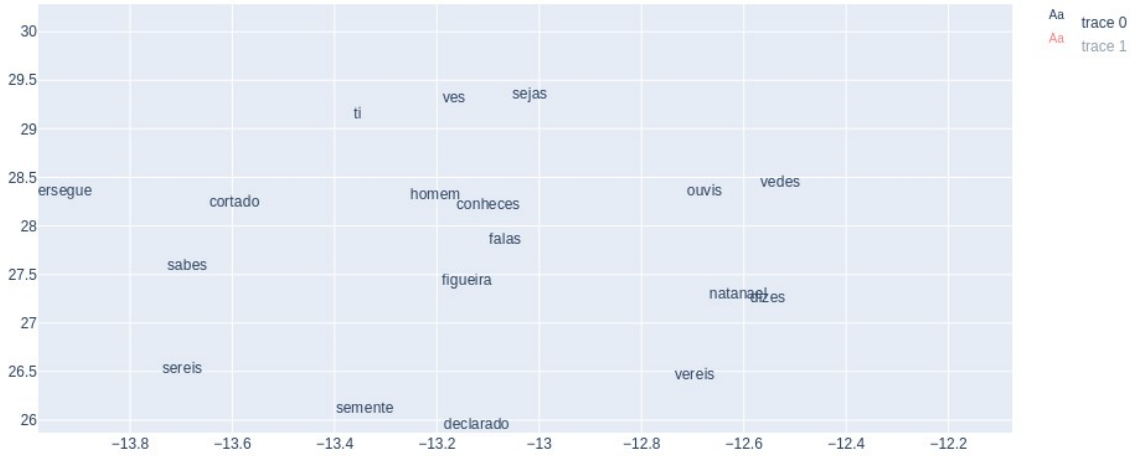
Para o período I vemos as palavras “conhecia”, “cuidados”, “sentes”, “comecara” mais próximas (Figura 14). Já para o período II vemos as palavras “conheces”, “falas”, “figueira”, “ouvis”, “ves”, “sejas” (Figura15). No Período III vê-se as palavras “miserável”, “ateu”, “ímpio”, “solene”, “livrar” (Figura 16).

Figura 14: Campo semântico da palavra "Homem", período I



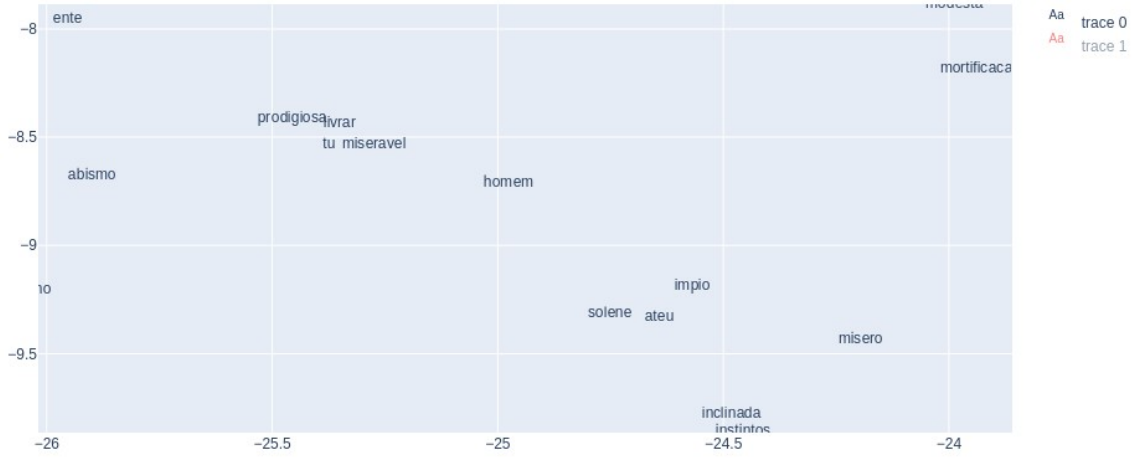
Fonte: Elaborado pelo autor.

Figura 15: Campo semântico da palavra "Homem", período II



Fonte: Elaborado pelo autor.

Figura 16: Campo semântico da palavra "Homem", período III



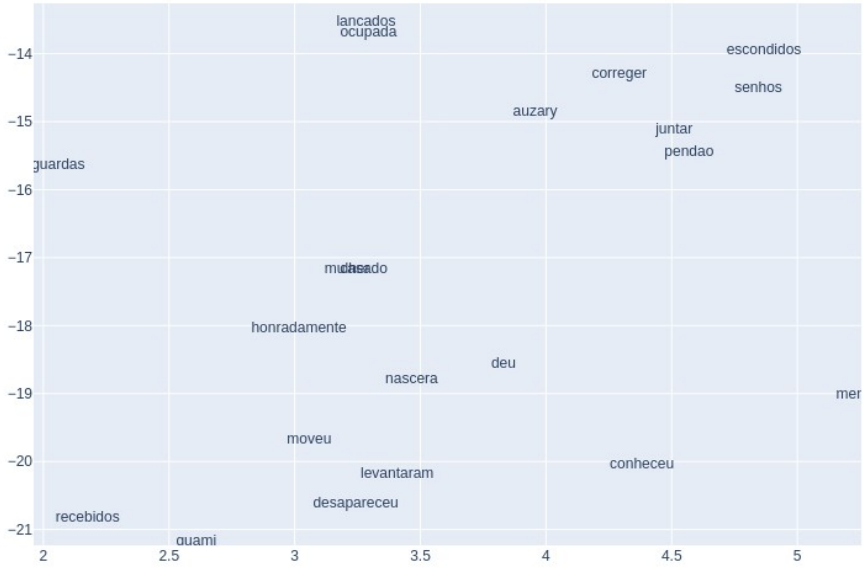
Fonte: Elaborado pelo autor

5.3 CAMPO SEMÂNTICOS PARA A PALAVRA “MULHER”

Para o campo semântico da palavra “Mulher”, vemos no período I que ela ficou bastante próxima da palavra “casado”, seguida das palavras “nascera”, “deu”, ‘nascera”, “honradamente” (Figura 17). Já no período II, temos as palavras “marido” e “adultério” muito próxi-

mas (Figura 18). Finalmente no período III, temos as palavras “desgraçada”, “coitadinha”, “marido” e “margarida” na proximidade da palavra analisada (Figura19).

Figura 17: Campo semântico da palavra "Mulher", período I

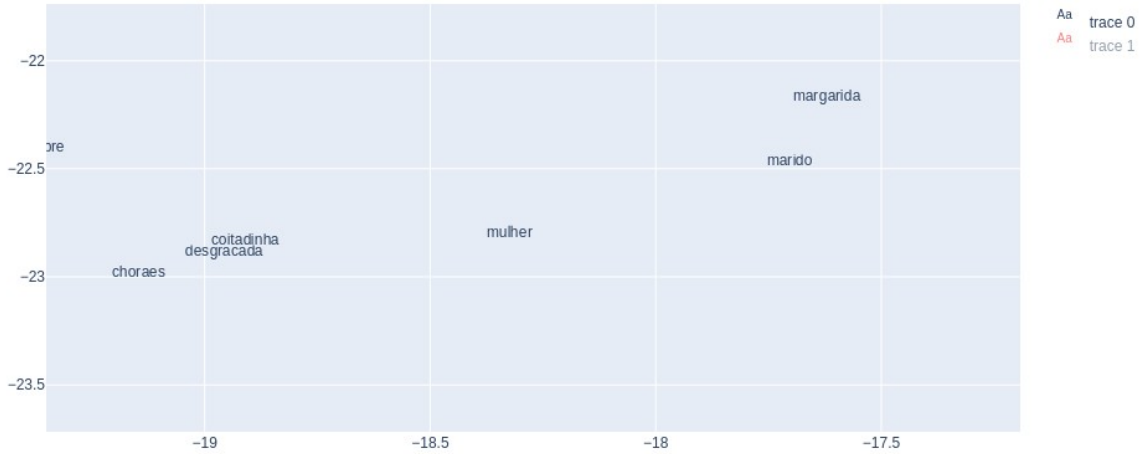


Fonte: Elaborado pelo autor.

Figura 18: Campo semântico da palavra "Mulher", período II

Fonte: Elaborado pelo autor.

Figura 19: Campo semântico da palavra "Mulher", período III



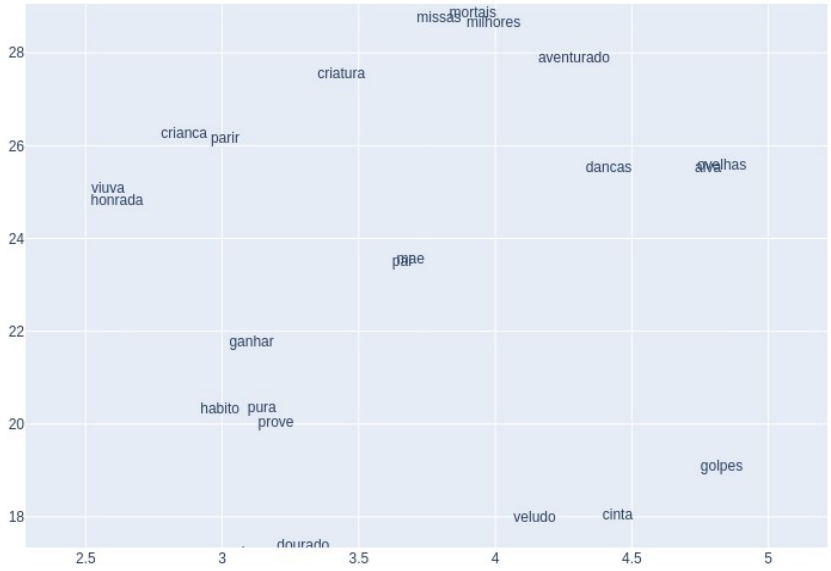
Fonte: Elaborado pelo autor.

5.4 CAMPOS SEMÂNTICOS PARA AS PALAVRAS “PAI” E “MÃE”

A análise do campo semântico dessas palavras foi realizada em conjunto devido a proximidade que elas se encontram nos resultados. Para o campo semântico no período II, elas se sobrepõe, como mostrado na figura 20. Temos as palavras “criatura”, “ganhar”, “cinta”, “pura”,“missas”, “parir”, “criança”, “viúva” na proximidade das palavras analisadas.

Já para o período II, as palavras foram analisadas separadamente. A palavra “Pai” possui em sua proximidade as palavras “testemunho”, “conheceis”, “credes”, “guardado”, “enviou” em sua vizinhança, como mostra a figura 21. Já a palavra “Mãe” se encontra próxima de expressões como “casados”, “bodas”, “legítimo”, “casal”, “embaracos” e “ajustada”.

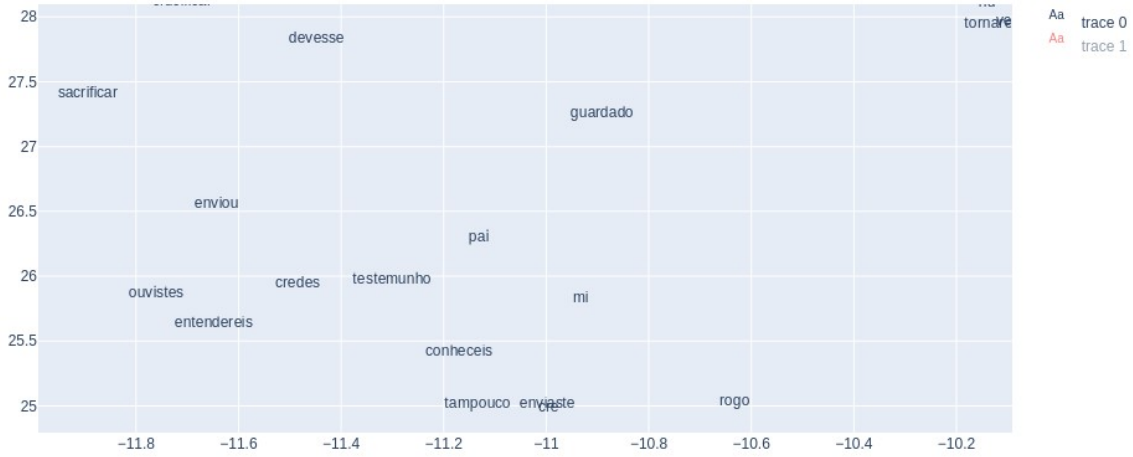
Figura 20: Campo semântico das palavras "pai" e "mãe", período I.



Fonte: Elaborado pelo autor.

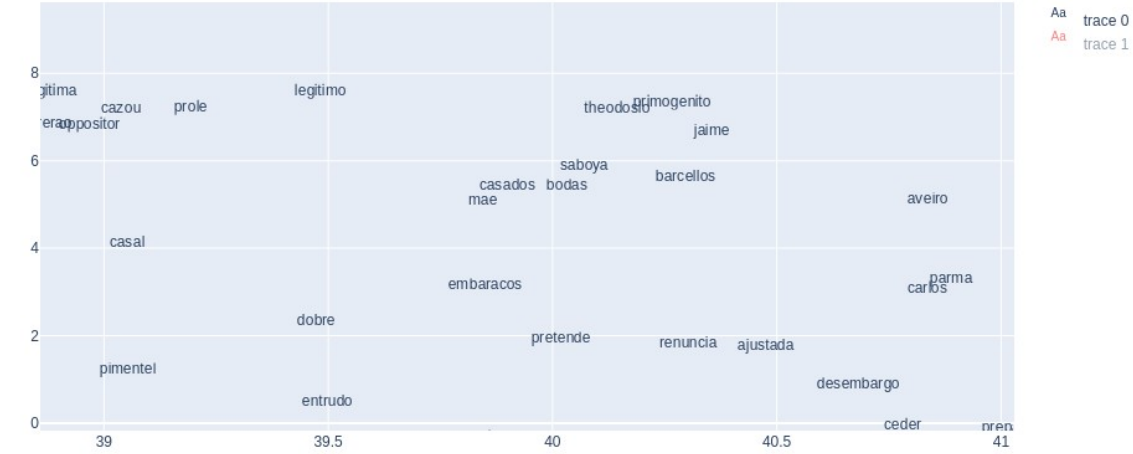
No período III, o campo semântico da palavra “Pai” mostra proximidade com “consolar”, “paterno”, “filho”, além de “continae” e “tourear”. Já o campo de “Mãe”, mostra as palavras “virtuosa”, “filha”, “carinhosa”.

Figura 21: Campo semântico da palavra "Pai", período II



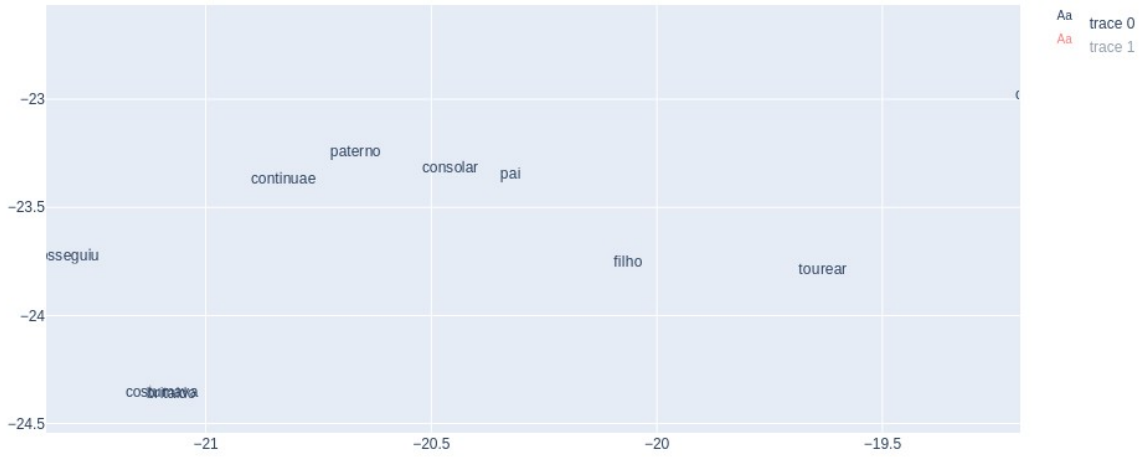
Fonte: Elaborado pelo autor

Figura 22: Campo semântico da palavra "Mãe", período II



Fonte: Elaborado pelo autor.

Figura 23: Campo semântico da palavra "Pai", período III



Fonte: Elaborado pelo autor.

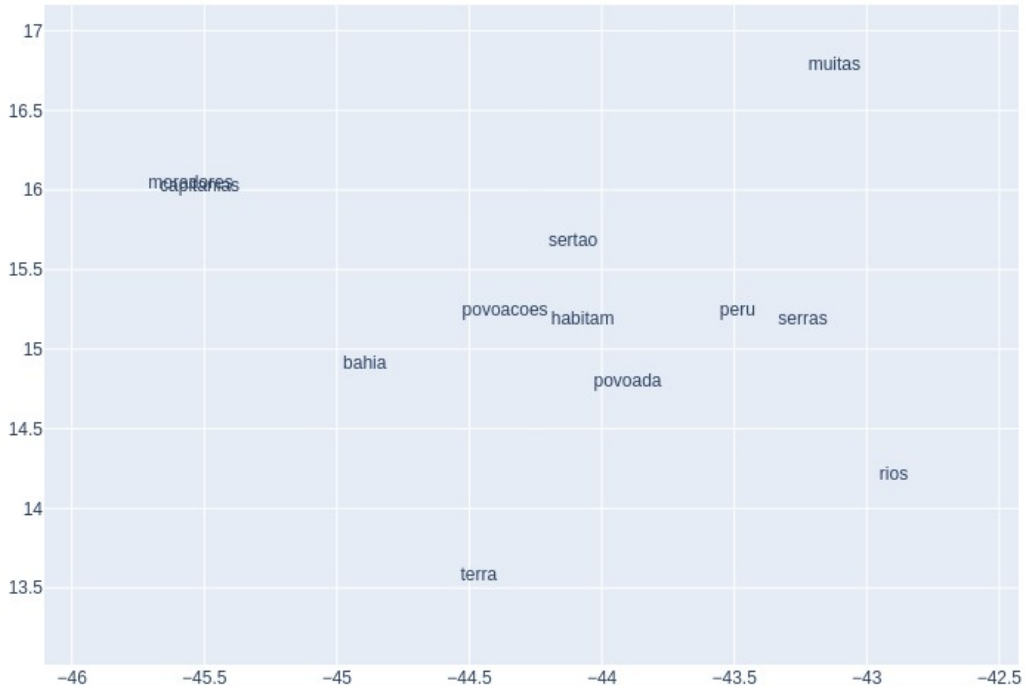
Figura 24: Campo semântico da palavra "Mãe", período III

Fonte: Elaborado pelo autor.

5.5 CAMPOS SEMÂNTICOS PARA A PALAVRA“TERRA”

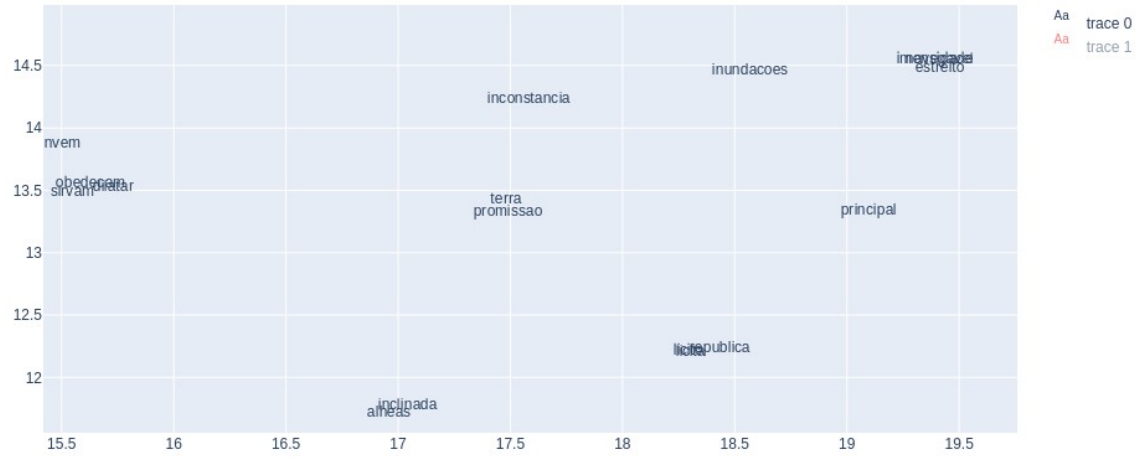
Para o período I, a palavra “Terra” possui em sua vizinhança as palavras “sertão”, “habitam”, “povoada”, “bahia”, “rios”, figura 25. Já para o período II a figura 26 mostra as palavras “inconstância”, “promissão” e destaca-se a palavra “republica”. Finalmente o período III apresenta palavras como “campinas”, “montanhas”, “ribeiras”, “ventos” e “tempestades, como visto na figura 27.

Figura 25: Campo semântico de "Terra", período I



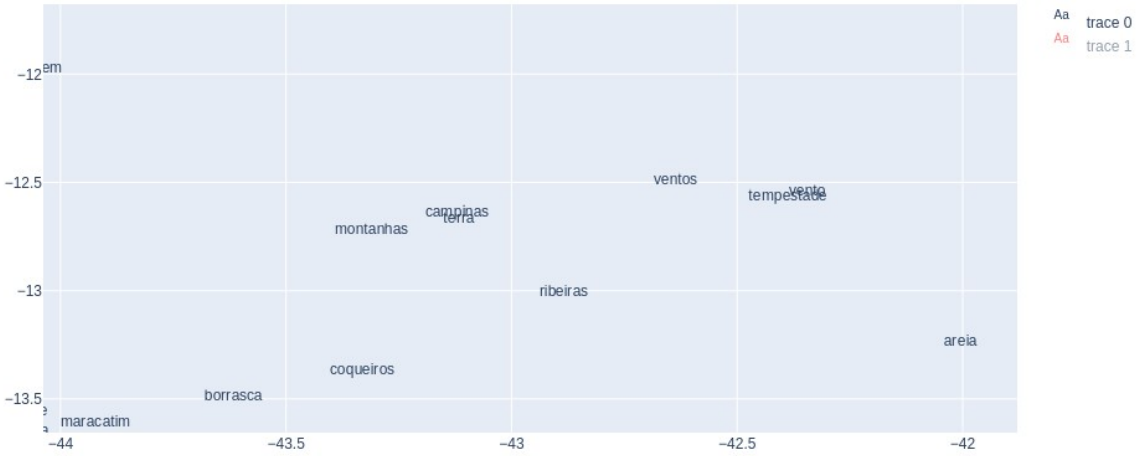
Fonte: Elaborado pelo autor

Figura 26: Campo semântico de "Terra", período II



Fonte: Elaborado pelo autor

Figura 27: Campo semântico para "Terra", período III



Fonte: Elaborado pelo autor

5.5 DISCUSSÃO DOS RESULTADOS

A primeira observação dos resultados é que a lematização não aconteceu de forma efetiva. Encontra-se formas verbais flexionadas, como “vedes”, “sejas”, “sabes”, “amas” (Figura e formas ajetivais apresentando o gênero feminino como em “coitadinha”, “desgraçada”.

Para a palavra “Deus” vê-se que ela surge em contextos ligados a religião. Isso surge como uma forma de validação para o modelo.

A palavra “Homem” não vem acompanhada, em um primeiro momento, de palavras que possam conferir significados associados a ela. Já no segundo período vemos a palavra associada a verbos sensoriais como “vedes”, “ouvis”, “falas”, e também ao verbo. Já no último momento vê-se a forma “homem” em palavras se viés negativo como “miserável”, “impio” e também a palavra “ateu” e a palavra com viés positivo “solene”.

A palavra “Mulher” aparece inicialmente em contexto não muito definido, com as palavras “honradamente”, “nascera”, “deu”, “moveu”. As poucas palavras encontradas para esse contexto podem se dar pela preferência por palavras como “rapariga” para se referenciar a mulher jovem. Já para o período II é interessante notar a relação próxima dos termos “marido”, “mulher” e “adultério”. Por fim, A palavra “mulher” continua próxima da palavra “marido”, mas também na vizinhança de termos como “coitadinha” e “desgraçada. Apesar da falha do lematizador em deflexionar essas expressões, elas apareceram próximas a palavra “mulher” mostrando a relação com o “gênero”.

Para as palavras “pai” e “mãe” temos inicialmente um bom resultado, pois aparecem bastante próximas campo semântico. Vê-se palavras relacionadas a família em suas proximidades como “criança”, “viúva” e também o verbo “parir” . Em seguida, para o segundo período, a palavra “pai” aparenta estar mais relacionada a um contexto religioso, como nas palavras “testemunho”, “credes”, “rogo” em sua proximidade. Já “mãe”, no mesmo período, mostra palavras relacionados a família e casamento como “casados”, “casal”, “prole”, “bodas”, “primogênito”. Por fim no período III, a palavra “pai” aparece ligada a palavras relacionadas ao contexto familiar como “paterno” e “filho”. Já a palavra “mãe” também aparece ligada a contextos familiares mostrando a palavra “filha” e “orfã” em sua proximidade, também vemos os adjetivos “virtuosa” e “carinhosa” próximos.

Por último a palavra “Terra” apresenta no período I, uma proximidade maior com termos relacionados a regiões brasileiras, como “bahia” e “sertão”. Já no período II a palavra parece ser encontrada em contextos diferentes, tendo em vista o surgimento da palavra “república”.

Os resultados apresentados mostram-se de qualidade variável. Em alguns casos, como a palavra “Terra”, apesar de apresentar apenas 2209 ocorrências no corpus, foi possível visualizar uma mudança no seu uso dentro dos três períodos analisados.

Já a forma “Deus” não apresentou variação notável em seu campo semântico. Apesar disso, o fato de essa forma estar sempre presa ao contexto religioso, surge como forma de validar o modelo, o que nem sempre é possível se fazer de forma quantitativa.

Por fim, a palavra “mulher” encontra-se associada a formas como “honrada”, “nasceu”, “coitadinha”, “desgraçada” e “adultério”, mostrando de certa forma as diferentes percepções ao longo do período.

Os processos de mudança (ou manutenção) semântica analisados anteriormente estão em conformidade com as propostas de Givón: palavras de sentido semelhantes foram agrupadas em regiões próximas nos modelos. Não foi possível entretanto verificar alguma mudança drástica de sentido, até porque essas palavras não sofrem necessariamente uma mudança de sentido ao longo do tempo. O que pode ser analisado, entretanto, é a vizinhança dessas palavras e, a partir disso, examinar como as formas estão organizadas no léxico disponível no corpus.

Em um caráter mais técnico, processo de lematização não foi eficiente devido a forma com que o processo é realizado. Para esse estudo utilizou-se a biblioteca Spacy, que possui um lematizador automático de acurácia relativamente baixa (76%). A lematização do corpus diacrônico sofreu também por possuir formas desconhecidas ao modelo, que utiliza um conjunto de regras para gerar os lemas.

Os resultados também foram limitados pelo tamanho o corpus e pel

6 CONCLUSÃO

O presente trabalho buscou analisar expressões significativas e os diferentes contextos semânticos onde elas surgem em diferentes períodos de tempo. A análise foi feita através do uso de técnicas de NLP como *Word Embeddings*, que permitem agrupar palavras de sentido próximas em um espaço vetorial e visualizar seus vizinhos.

Destaca-se aqui a importância do processo de lematização, que permite agrupar palavras flexionadas dentro da mesma forma, tornando assim a análise mais próxima. Esse processo é de extrema importância para línguas de morfologia rica, como o Português, e não recebe tanta atenção devido aos sistemas de NLP serem desenvolvidos principalmente para o inglês, que possui morfologia pobre.

Os obstáculos para um melhor resultado se dão principalmente pelo tamanho do corpus. Os resultados apresentados por Hamilton et al (2016) foram obtidos através de um corpus que possui mais de 100 milhões de tokens por período, um valor muito distante da quantidadae de tokens obtidos com o corpus Tycho Brahe. Apesar de os autores citarem outras formas de obtenção de *Word Embeddings*, e também recomendar seu uso para corpora menores, não há uma forma definitiva de se determinar o que é um corpus “pequeno” ou “grande”, sendo esse conceito determinado pela tarefa a se realizar.

Por fim, o caráter misto do corpus também influenciou os resultados. O corpus Tycho Brahe possui tanto textos de cartas, poemas e peças de teatro, como textos jornalísticos. Ou seja, não é limitado a apenas um gênero textual, o que influencia nas mudanças detectadas.

Considerando as limitações encontradas nessa análise, propõe-se para estudos futuros o uso de formas alternativas de obtenção de *Word Embeddings*, comparando os resultados deste com os aqui obtidos. Além disso, sugere-se o uso de outro lematizador, que possa fornecer um resultado satisfatório. Pode-se também realizar novos recortes temporais e analisá-los a fim de buscar diferentes relações semânticas.

Todos os desafios citados anteriormente decorrem do caráter pioneiro do presente estudo para o Português, não existindo uma análise quantitativa que utilize metodologia similar para dados diacrônicos. Assim, o caráter exploratório deve ser levado em consideração ao se analisa os méritos deste trabalho.

7 REFERÊNCIAS

BECHARA, E. **As fases históricas da língua portuguesa: tentativa de proposta de nova periodização**. [s.l.] Universidade Federal Fluminense., 1985.

CAMBRAIA, C. N. Da lexicologia social a uma lexicologia sócio-histórica: caminhos possíveis. **Revista de Estudos da Linguagem**, v. 21, n. 1, p. 157–188, 2013.

CUNHA, A. F. DA. Funcionalismo. **Manual de linguística. São Paulo: contexto**, v. 2, p. 157–176, 2008.

DE SOUSA, M. C. P. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. **Filologia e Linguística Portuguesa**, v. 16, n. esp., p. 53–93, 2014.

GIVÓN, T. **Functionalism and grammar**. [s.l.] John Benjamins Publishing, 1995.

GIVÓN, T. **Syntax: an introduction**. [s.l.] John Benjamins Publishing, 2001. v. 1

HAMILTON, W. L.; LESKOVEC, J.; JURAFSKY, D. Diachronic word embeddings reveal statistical laws of semantic change. **arXiv preprint arXiv:1605.09096**, 2016.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing. **Upper Saddle River, NJ: Prentice Hall**, 2008.

MICHEL, J.-B. et al. Quantitative analysis of culture using millions of digitized books. **science**, v. 331, n. 6014, p. 176–182, 2011.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

ROBIN, R.; DE MENESES BOLLE, A. B. **História e lingüística**. [s.l.] Editora Cultrix, 1977.

SARDINHA, T. B. **Lingüística de corpus**. [s.l.] Editora Manole Ltda, 2004.

SWINNEY, D. A. Lexical access during sentence comprehension:(Re) consideration of context effects. **Journal of verbal learning and verbal behavior**, v. 18, n. 6, p. 645–659, 1979.

tmikolov/word2vec: Automatically exported from code.google.com/p/word2vec. Disponível em: <https://github.com/tmikolov/word2vec>. Acesso em: 16 ago. 2021.