UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
BACHARELADO EM ESTUDOS LINGUÍSTICOS

Lucas Fonseca Lage

Mudança semântica e word embeddings: estudos de caso na diacronia do português

Belo Horizonte
2021

Lucas Fonseca Lage

Mudança semântica e word embeddings: estudos de caso na diacronia do português

Trabalho de 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 Estudos Linguísticos
Orientador: Prof. Dr. Evandro Landulfo Teixeira Paradela Cunha

Belo Horizonte

2021

Lucas Fonseca Lage

Mudança Semântica e Word Embeddings: estudos de caso na diacronia do Português

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de bacharel em Estudos Linguísticos e aprovado em sua forma final

Belo Horizonte, 03 de Setembro de 2021.

Banca Examinadora:

Prof. Dr. Evandro Landulfo Teixeira Paradela Cunha,
Orientador
Universidade Federal de Minas Gerais

Prof.ª Dr.ª Adriana Silvina Pagano,
Universidade Federal de Minas Gerais

Prof. Dr. Bruno Neves Rati de Melo Rocha,
Universidade Federal de Minas Gerais

The great art of the historical linguist is to make the best of this bad data - “bad” in the sense that it may be fragmentary, corrupted, or many times removed from the actual productions of native speakers (LABOV, 1972).

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. Este trabalho se propõe analisar os contextos de ocorrência das palavras “deus”, “homem”, “mulher”, “pai”, “mae” e “terra” no corpus Tycho Brahe do Português. Para isso utiliza-se o algoritmo Skip-gram para gerar os Word Embeddings, e, posteriormente serem geradas visualizações para a rede de relações semânticas de cada palavra em três diferentes recorte temporais.

Palavras-chave: Linguística Computacional. Estudos Diacrônicos. Processamento de Língua Natural. Mudança Linguística. Vetorização de Palavras.

ABSTRACT

According to Givón (2001) the lexicon is a repository of relatively time-stable, socially shared and well-coded concepts, which are also organized in a network-like structure, where similar concepts are grouped together. On a similar note, lexicologist Georges Matoré states that words define associative connections between themselves and also proposes the notional field and testimony words, elements which organize the lexicon. Using computational tools such as word embeddings, that allow words to represented as vectors in a vector space, it is possible to analyze words grouped together by same semantic traces. This thesis analyses contexts in which the following words occur: “deus”, “homem”, “mulher”, “pai”, “mae” and “terra” in the Tycho Brahe corpus for Portuguese. The Skip-gram algorithm was used to create word embeddings and plot graphics for each word and their semantic relationship network in three different periods of time.

Keywords: Computational Linguistics. Diachronic Studies. Natural Language Processing. Linguistic Change. Word Embeddings.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Muitas são as propostas a respeito do que causa a mudança de significado. Givón, funcionalista, em seus trabalhos afirma que o léxico é um repositório de conceitos relativamente estáveis no tempo, compartilhados socialmente e bem codificados. Além disso, esses conceitos são interconectados em uma rede, onde a ativação de um conceito leva a ativação de um conceito vizinho (GIVÓN, 1995).

De acordo com o autor a língua humana evoluiu em paralelo com mecanismos cognitivos, da organização sócio cultural e das habilidades comunicativas de hominídeos. Assim, em uma sociedade onde evoluções tecnológicas e culturais são a norma, a possibilidade de transmitir conhecimento e habilidades é de grande valia, mas com a variância de conceitos relevantes socialmente, algumas formas tornam-se mais frequentes e outras caem em desuso (GIVÓN, 1995).

Ao analisar essas formas em uso e as formas associadas a elas,o lexicógrafo Georges Matoré apresenta os conceitos de *palavra testemunho* e *campo nocional*, que seriam usados para descrever termos social e culturalmente relevantes. Matoré chega a apresentar redes associativas para certo termos e busca também analisar como essas redes se alteram ao longo do tempo. Entretanto, sua metodologia foi bastante criticada e os estudos lexicológicos sociais foram negligenciados (CAMBRAIA, 2013).

Tendo em vista os recentes desenvolvimentos de técnicas na área da computação, em especial as técnicas de *word embeddings* (vetorização de palavras), novos tipos de análise linguística têm 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 e censura. Não só isso, mas essas mesmas medidas simples podem auxiliar tanto na decisão de inclusão de novos termos em um dicionário quanto na retirada de termos irrelevantes (MICHEL et al., 2011).

A frequência de uso de uma palavra por si só é um dado interessante, mas ela também se altera por diversos motivos. Givón, em sua perspectiva Funcionalista, afirma que a forma deve cumprir uma função, logo quando a função se torna pouco útil, a forma cai em desuso. Similarmente, Bochkarev, Solovyev e Wichmann (2014) afirmam que a mudança lexical é favorecida não só

por mudanças em um ambiente social e natural, como também em ambiente linguístico particular. Esses autores utilizam um corpus diacrônico para analisar a taxa de mudança lexical de acordo com a frequência de ocorrência das 100.000 palavras mais frequentes, e mostram como o inglês britânico e o inglês americano estão convergindo, apesar da inicial separação. Eles afirmam que os dois idiomas se tornaram mais próximos dada ao advento da mídia de massa, que cresceu exponencialmente nos séculos XX e XXI (GIVÓN, 2001; BOCHKAREV; SOLOVYEV; WICHMANN, 2014).

Já em um caráter mais específico, Hamilton, Leskovec e Jurafsky, (2016) usam técnicas de vetorização de palavras não só para buscar palavras que sofreram mudança de significado, mas também para validar mudanças de significado já conhecidas. Com sua análise, é possível também buscar as formas que passaram por maiores mudanças semânticas. Após o uso dessas técnicas e com a validação de seus próprios trabalhos os autores se sentem confortáveis em apresentar leis para a mudança semântica (HAMILTON; LESKOVEC; JURAFSKY, 2016).

Buscando otimizar o trabalho de Hamilton, Leskovec e Jurafsky (2016), foi proposto por Yao et al. (2018) uma nova metodologia de aprendizado de vetores de palavras capaz de codificar também o componente tempo. Esses autores, entretanto, não analisam o significado de formas específicas, mas as palavras associadas a elas. Assim, o algoritmo desenvolvido é capaz de codificar associações como “apple” e “strawberry” em um primeiro momento, e em outro, “apple” e “iphone” (YAO et al., 2018).

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 Língua Natural (NLP na sigla em inglês). A princípio utiliza-se da análise das frequências de ocorrências das palavras de um corpus diacrônico do português, em seguida são gerados vetores de palavras para três recortes temporais distintos desse corpus e, finalmente, são geradas imagens onde se é possível visualizar as redes de relações semânticas ao longo do tempo.

A fundamentação teórica para esta pesquisa se pautará em uma abordagem lexicológica e funcional, tendo em vista que os conceitos aplicados de forma prática pelos algoritmos de *Word Embeddings* foram parcialmente fundamentados teoricamente muito antes por autores como Georges Matoré e Talmy Givón (GIVÓN, 1995; CAMBRAIA, 2013).

Como fonte de dados é utilizado o corpus diacrônico do português brasileiro Corpus Tycho Brahe, organizado pela Unicamp, que abrange textos do século XIII até o século XX. (DE SOUSA, 2014).

1.1 OBJETIVOS

O trabalho visa, de maneira geral, realizar uma análise de mudança semântica diacrônica. 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 Língua Natural.

São então os objetivos específicos:

- Analisar mudanças semânticas com o auxílio de técnicas de NLP;
- Avaliar a viabilidade de técnicas recentes de computação em estudos linguísticos;
- Gerar métodos capazes de visualizar graficamente redes de relações de sentido.

2 REVISÃO BIBLIOGRÁFICA

2.1 GIVÓN E O FUNCIONALISMO

Segundo Cunha (2008), 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 nos métodos utilizados, nos dados considerados relevantes e, mais profundamente, nos objetivos da análise linguística (CUNHA, 2008).

Dentro de uma perspectiva funcionalista, uma sentença é 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 generalizações. 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 influenciados por essas conclusões não foi ainda estudada (GIVÓN, 1995, 2001).

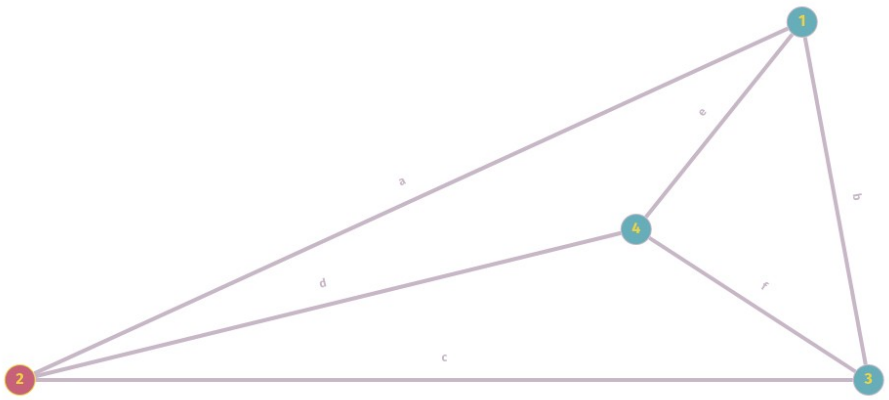
Em seu livro *Functionalism and Grammar*, o autor lista os componentes funcionais para a comunicação humana e os divide em dois módulos principais, que interagem entre si. São eles o *Sistema de Representação Cognitiva* e os *Sistemas de Codificação*. Dentro do Sistema de Representação Cognitiva temos 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 às 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 sociocultural, e ao nosso universo mental-interno. Além disso, os conceitos que compõem 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 possuirá o mesmo significado 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 arestas, e que cada nó corresponde a uma palavra. A ativação de um nó-palavra ainda seria responsável pela ativação de outros nós-palavra que possuam uma relação íntima com o primeiro. Na figura 1 tem-se um exemplo da rede descrita por Givón, que se assemelha a um grafo, nela vemos os nós de 1 a 4 e as arestas *a* a *f*. 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).

Figura 1: Exemplo de grafo.



Fonte: Elaborado pelo autor.

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 trabalhou com o conceito do léxico em rede buscando uma lexicologia social.

2.2 LÉXICO EM REDE

Através de seu livro *La lexicologie sociale*, Georges Matoré cita a influência de fatores sociais no estudo do léxico. Ele propõe uma série de princípios para 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 à 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).

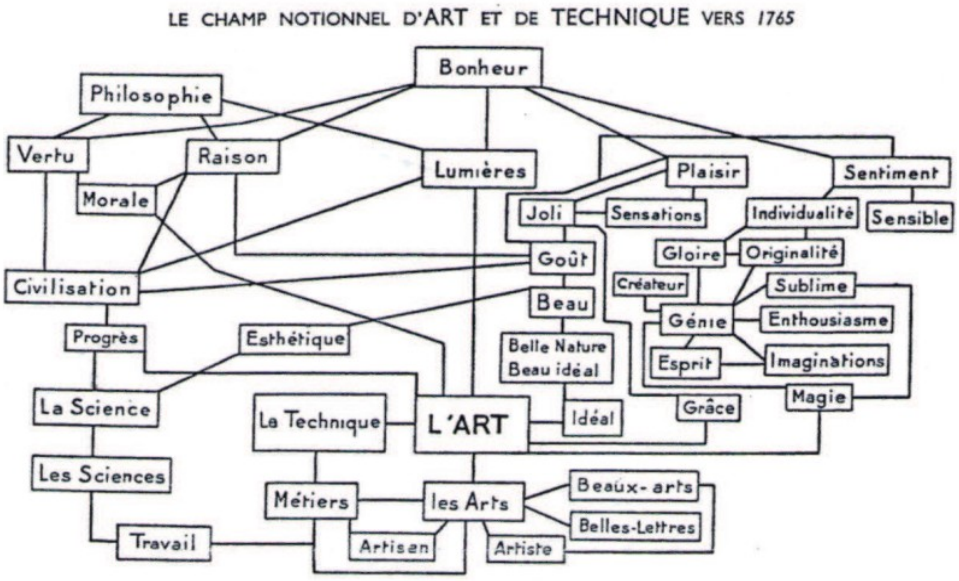
A interpretação de Matoré sobre como o vocabulário se comporta é sistêmica, ou seja, admite que as palavras estabelecem relações recíprocas na consciência. As palavras podem se relacionar com suas vizinhas, através de relações sintagmáticas, ou com palavras similares, através de forma ou sentido, estabelecendo 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).

São encontradas algumas similaridades entre o trabalho de Matoré e os princípios estabelecidos por Saussure, para o estruturalismo, mas o lexicólogo se desvencilha dessa corrente. Ele o faz discordando de Saussure a respeito da organização morfológica do léxico e atribui ao fator social o principal papel na organização do léxico (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, de-

vem 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 2, e o campo nocional de Artista entre os anos de 1827 e 1834, Figura 3 (CAMBRAIA, 2013).

Figura 2: Campo nocional de "Arte" e "Técnica" em 1765, segundo Georges Matoré

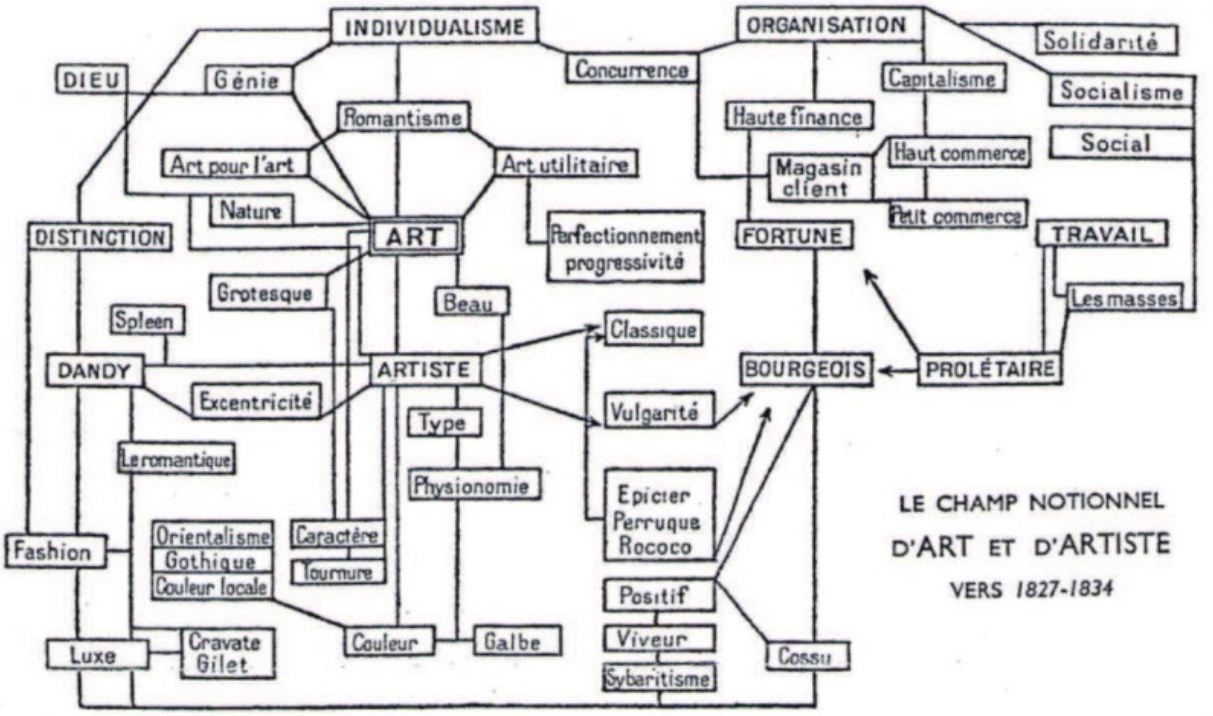


Fonte: Cambraia, 2013

Os métodos defendidos por Matoré foram muito criticados na 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, ao criticar os métodos, que os estudos de Matoré não poderiam refletir a sociedade como um todo, mas apenas os 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).

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. (CAMBRAIA, 2013).

Figura 3: Campo nocional de "Artista" entre 1827-1834, segundo Georges Matoré.



Fonte: Cambraia, 2013

Cambraia (2013) apresenta um estudo de caso, levando em consideração elementos metodológicos que deve constituir uma lexicologia sócio-histórica. Entre esses elementos temos a articulação entre fatores intralinguísticos e extralinguísticos, que, até naquele momento, não era considerada em estudos do léxico. Para isso o autor apresenta um estudo das expressões “esquadrão da morte” e “grupo de extermínio” em jornais brasileiros, e analisa a frequência de ocorrências ao longo do tempo e os contextos em que aparecem. Assim, o autor propõe que um modelo de lexicologia sócio-histórica deve pressupor estudos de casos específicos, que possam ser relacionados entre si com base em dados sócio-históricos (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 trabalhos pautados nessa perspectiva carecem de uma metodologia que pudesse capturar essas complexidades.

2.3 WORD EMBEDDINGS

2.3.1 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 como Firth quando eles perceberam que palavras sinônimas ocorrem no mesmo ambiente. Firth afirma que “Você conhece uma palavra pela sua companhia.” (FIRTH, 1957; JURAFSKY; MARTIN, 2008).

Enquanto muitas palavras não possuem sinônimos, a grande maioria das palavras possui outras que são muito similares. Por exemplo, apesar de as 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 similaridade entre palavras, sentenças ou documentos é bastante útil em diversas tarefas de NLP 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 e colegas, onde são criados três eixos a fim de avaliar o sentido afetivo de uma palavra e então associa-se um valor numérico a cada eixo. Na Figura 4 vemos as valorações das palavras *“courageous”*, *“music”*, *“heartbreak”* e *“cub”* em três eixos. O eixo de valência está relacionado a agradabilidade do estímulo gerado, o de excitação diz respeito a intensidade da emoção provocada pelo estímulo e a dimensão de dominância se refere ao grau de controle exercido pelo estímulo (OSGOOD; SUCI; TANNENBAUM, 1957; JURAFSKY; MARTIN, 2008).

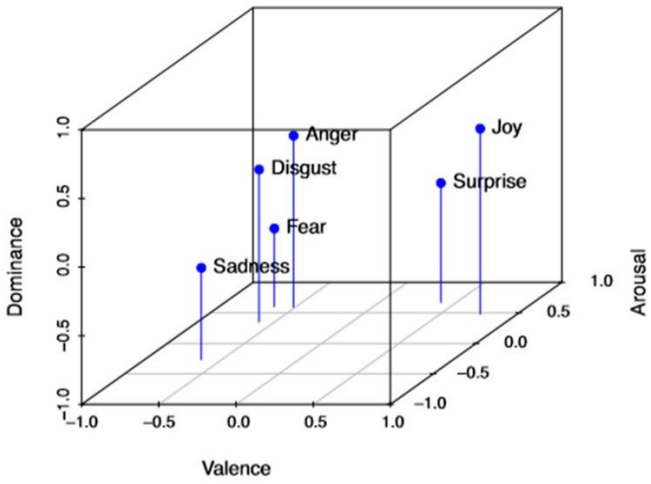
Figura 4: Variação quantitativa do sentido afetivo em três eixos

	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. A Figura 5 mostra um exemplo de como palavras relacionadas a sentimentos estão posicionadas nesse espaço vetorial. As palavras mostradas são “anger”, “disgust”, “fear”, “joy”, “sadness” e “surprise”, ou “raiva”, “repulsa”, “medo”, “alegria”, “tristeza” e “surpresa” para o português.

Figura 5: Representação afetiva de palavras, segundo Osgood et al.



Fonte: Bălan et al., 2020

Após esses estudos iniciais, os métodos para obtenção de representações vetoriais evoluíram até chegarem a usar métodos de aprendizado automático, como os modelos *continous bag-of-words* e *Skip-gram*.

2.3.2 Continuous Bag-of-words e Skip-Gram

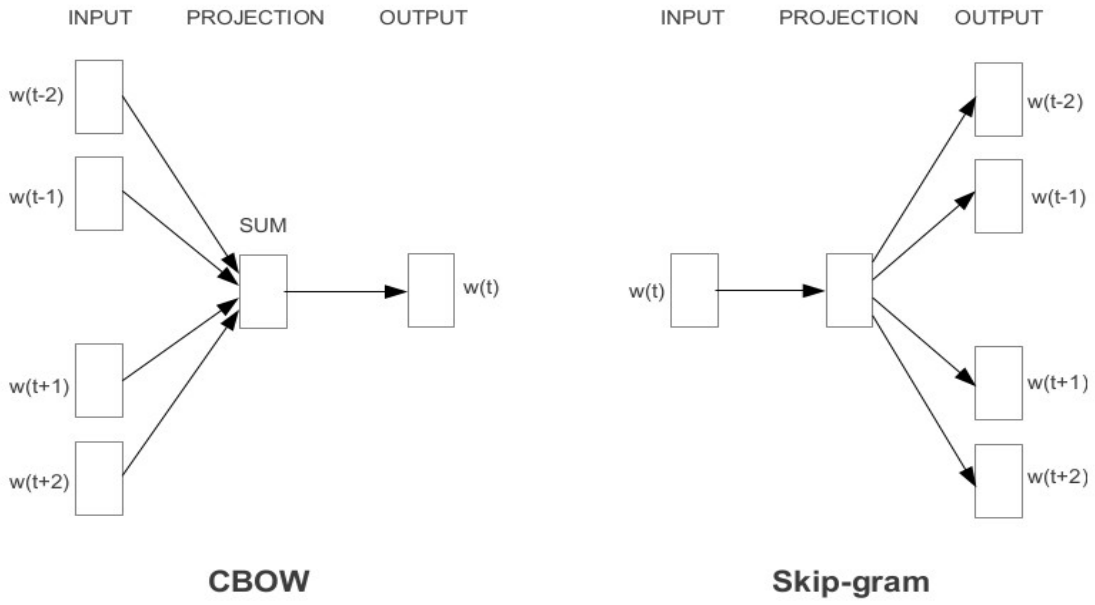
Um dos principais trabalhos com *Word Embeddings* é a publicação de Mikolov et al. (2013). Nele é proposta uma arquitetura de redes neurais capaz de criar uma representação vetorial

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 *Continuous 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, isto é, prediz o contexto a partir de uma palavra (MIKOLOV et al., 2013).

Figura 6: Esquema da arquitetura dos modelos CBOW e Skip-gram



Fonte: Mikolov et al., 2013

Para a 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, Mikolov et al. criaram testes 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". Essa sentença pode ser traduzida em uma operação vetorial como: Vetor("Paris") + Vetor("França") - Vetor("Alemanha") = vetor("Berlin") (MIKOLOV et al., 2013).

Dentro dessas tarefas, o modelo Skip-gram atingiu a melhor acurácia total quando comparado ao modelo CBOW e outras arquiteturas semelhantes. 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 finalizarem o treinamento, enquanto as outras arquiteturas precisaram de 14 dias de treino para criar o seu modelo (MIKOLOV et al., 2013).

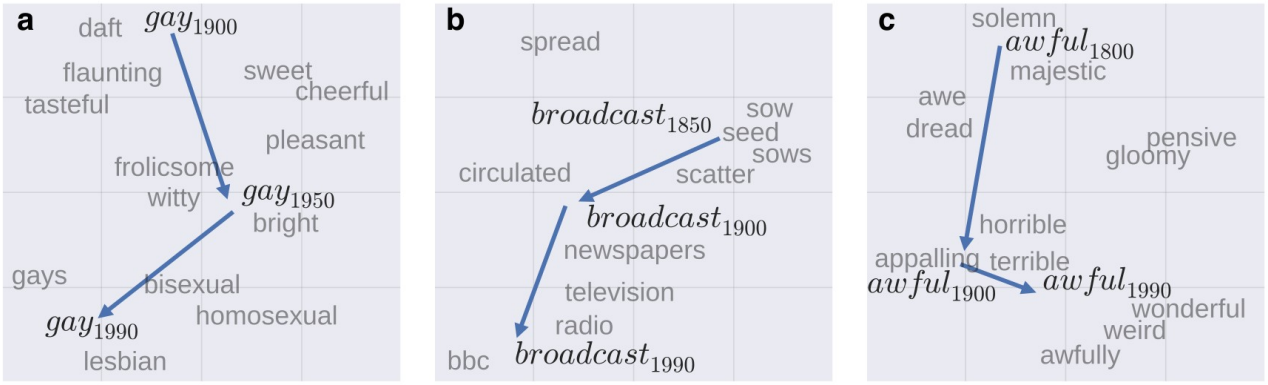
Tendo em mãos uma técnica não só capaz de capturar relações sintáticas e semânticas, como também de menor custo computacional, falta então a capacidade de se trabalhar com um corpus diacrônico.

2.4 WORD EMBEDDINGS E CORPORA DIACRÔNICOS

Hamilton, Leskovec e Jurafsky (2016) propõem 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 7 (HAMILTON; LESKOVEC; JURAFSKY, 2016).

Figura 7: Deslocamento vetorial das formas *broadcast*, *gay* e *awful* entre 1800 e 1990.



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 rádio (HAMILTON; LESKOVEC; JURAFSKY, 2016).

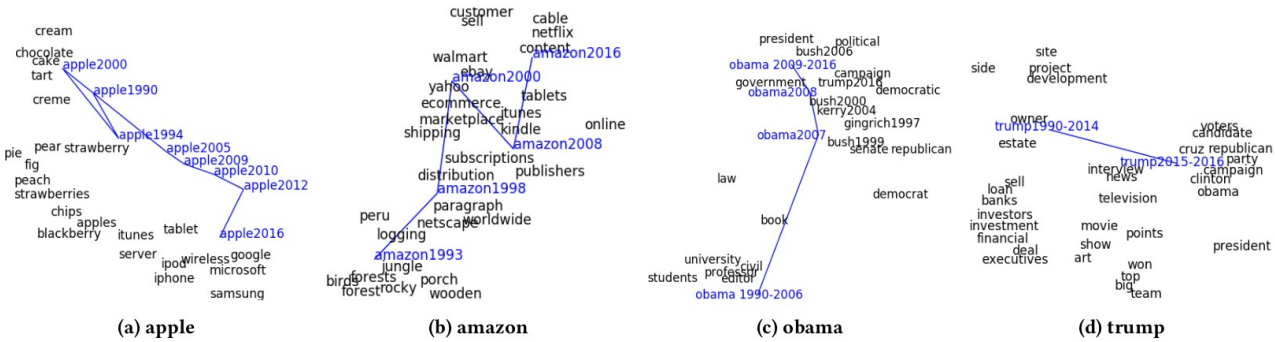
O estudo conseguiu, em um primeiro momento, validar mudanças de significado já conhecidas e foi em seguida usado para buscar as palavras que sofreram maior mudança semântica, aqui considerado o maior deslocamento no espaço vetorial ao longo dos períodos analisados (JURAFSKY; MARTIN, 2008).

Buscando otimizar a visualização de vetores em corpora diacrônicos, Yao et al. (2018) apresentam um novo modelo capaz de aprender vetores de palavras diacrônicos em um único passo. O maior problema ao se realizar esse tipo de vetorização em corpora diacrônico é alinhar os eixos dos modelos. Devido as operações realizadas nos vetores para a obtenção de um modelo como o Skip-gram, as mesmas palavas podem ser geradas em pontos diferentes do espaço vetorial. Isso não altera a distância entre elas dentro do mesmo modelo, mas entre modelos diferentes não é possível

analisar-se o deslocamento da palavra no intervalo de tempo entre os dois modelos. O método proposto por Hamilton, Leskovec e Jurafsky (2016) é constituído de dois passos: primeiro cria-se os vetores de palavras, em seguida alinha-se esses vetores em um mesmo eixo. A proposta de Yao et al. (2018) busca realizar a codificação do fator tempo paralelamente ao treinamento do modelo (HAMILTON; LESKOVEC; JURAFSKY, 2016; YAO et al., 2018).

Por fim, os autores apresentam redes associativas temporais para as palavras “apple”, “amazon”, “obama” e “trump”, vistos na Figura 8.

Figura 8: Trajetórias de nomes através do tempo, de acordo com Yao et al.



Fonte: Yao et al, (2018)

As associações aprendidas mostram como, por exemplo, o termo “amazon”, inicialmente associado a termos do campo da natureza, se torna associado a termos do campo da tecnologia. Assim, mostra-se que é possível analisar diacronicamente e a partir de métodos computacionais, as redes de associações dentro do léxico.

3 METODOLOGIA

A partir da visão de Givón sobre o léxico e da 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).

Buscando uma metodologia similar à de Hamilton, Leskovec e Jurafsky (2016), o corpus foi divido em períodos relevantes para o estudo da 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 (em torno de 8 milhões de tokens) é muito menor quando comparada aos corpora utilizados por Hamilton, Leskovec e Jurafsky (2016), que possuem 850 bilhões de tokens. 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 foram considerados apenas textos a partir do século XV (HAMILTON; LESKOVEC; JURAFSKY, 2016).

Primeiramente foi realizada uma análise exploratória a fim de encontrar as formas mais frequentes, a distribuição das frequências das palavras e quais delas são relevantes social e culturalmente.

As palavras foram escolhidas e analisadas através da perspectiva de *palavras-testemunho* e *campos nocionais* de Matoré. Dessa forma, buscou-se as redes de relações para os seguintes termos: *homem, mulher, pai, mãe, terra deus*. As palavras aqui mencionadas foram escolhidas por dois motivos: primeiramente são centrais para articulação de valores sócio culturais, são termos geralmente carregados e podem evidenciar vieses; segundo, possuem frequência relativamente alta dentro do corpus e possibilita uma melhor qualidade de resultados. A partir daí foram buscadas associações que possam dar pistas sobre a percepção e os conceitos culturais que permearam esses termos durante a história da língua portuguesa.

As análises foram desenvolvidas utilizando a linguagem de programação Python, além das diversas bibliotecas para NLP existentes como Spacy e NLTK (Natural Language Toolkit).

3.1 DESCRIÇÃO DO CORPUS TYCHO BRAHE

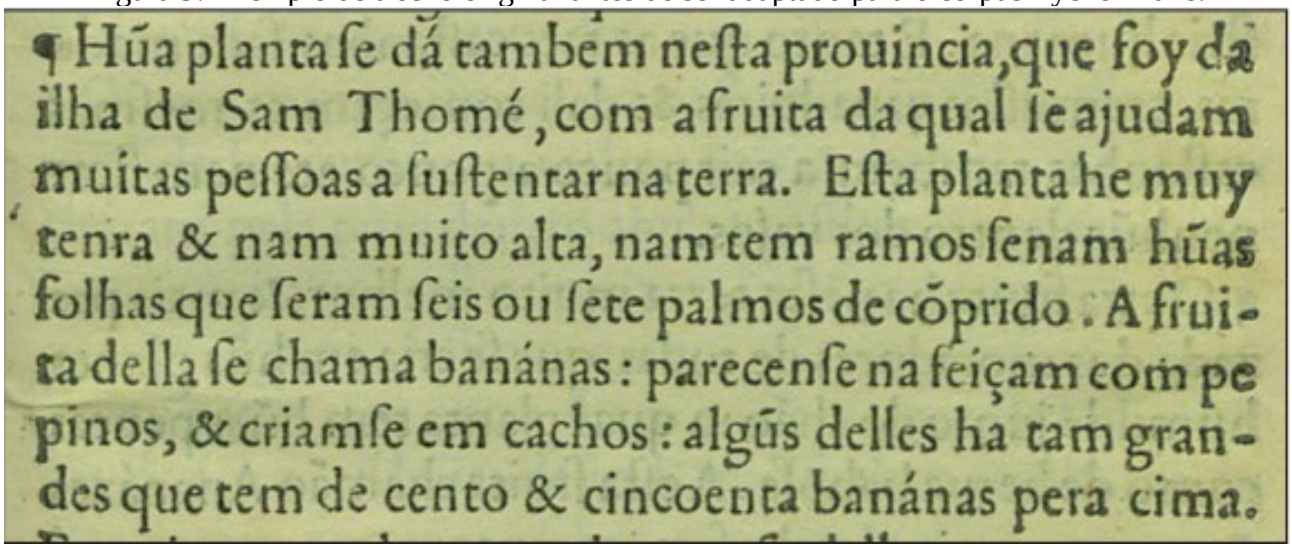
O Corpus Anotado do Português Histórico Tycho Brahe (CTB) foi pioneiro no que concerne a língua portuguesa, e permanece hoje como o maior corpus eletrônico anotado de textos históricos em português. Hoje, o conjunto de dados inclui textos escritos por autores portugueses, brasileiros e africanos, nascidos entre 1380 e 1845, publicados entre os século XIV e XX. 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, sendo 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 provêm de séculos passados, o seu processamento 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 três 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, expansão de abreviaturas, atualização ortográfica. A transcrição é feita com o auxílio da ferramenta *e-Dictor*, que permite a realização de adequações de grafia. Por exemplo, o termo original *parecenſe* é normalizado grafematicamente para *parecense*, e em seguida tem a grafia atualizada para *parecem-se*. Na Figura 9 temos um exemplo de como os textos originais se encontram (DE SOUSA, 2014).

Figura 9: Exemplo de trecho original antes de ser adaptado para o 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%. 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.). Um exemplo de frase anotada com esse anotador pode ser visto na figura10.

Figura 10: 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. Um exemplo do uso do parser pode ser visto na seguinte Figura 11 (DE SOUSA, 2014).

Figura 11: 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 foi treinado ao ser realimentado com seus resultados corrigidos por pesquisadores até que seu desempenho foi considerado satisfatório (DE SOUSA, 2014).

Dadas essas diversas características do corpus Tycho Brahe, faz-se então uma análise exploratória preliminar do corpus.

3.2 ANÁLISE EXPLORATÓRIA

A análise exploratória foi realizada com a intenção de se avaliar quais termos atenderiam as exigências do trabalho. As palavras a serem analisadas devem possuir ocorrência significativa para serem gerados vetores de qualidade e, também, serem relevantes social e culturalmente, possibilitando uma análise de seus contextos de uso e de formas relacionadas.

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

- Tokenização;
- Padronização em caixa baixa;
- Remoção de s*topwords*;
- Remoção de acentuação.

A *tokenização* consiste segmentar o texto em pedaços menores que possuam relevância para a análise, esses pedaços podem ser frases, palavras, símbolos gráficos ou numerais, desde que sejam relevantes. O resultado da *tokenização* é o *token*, aqui, refere-se ao nível da palavra.

Em seguida o texto é padronizado completamente em caixa baixa, evitando assim que *tokens* iguais sejam considerados diferentes devido a escrita em maiúscula. Dessa forma, os *tokens* “Portanto” e “portanto” correspondem ao mesmo *token* “portanto”.

A retirada de s*topwords* é um processo comum em tarefas de NLP. As *stopwords* são palavras consideradas gramaticais ou pouco relevantes semanticamente, como pronomes, preposições e artigos. 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 (BIRD; KLEIN; LOPER, 2009).

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

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

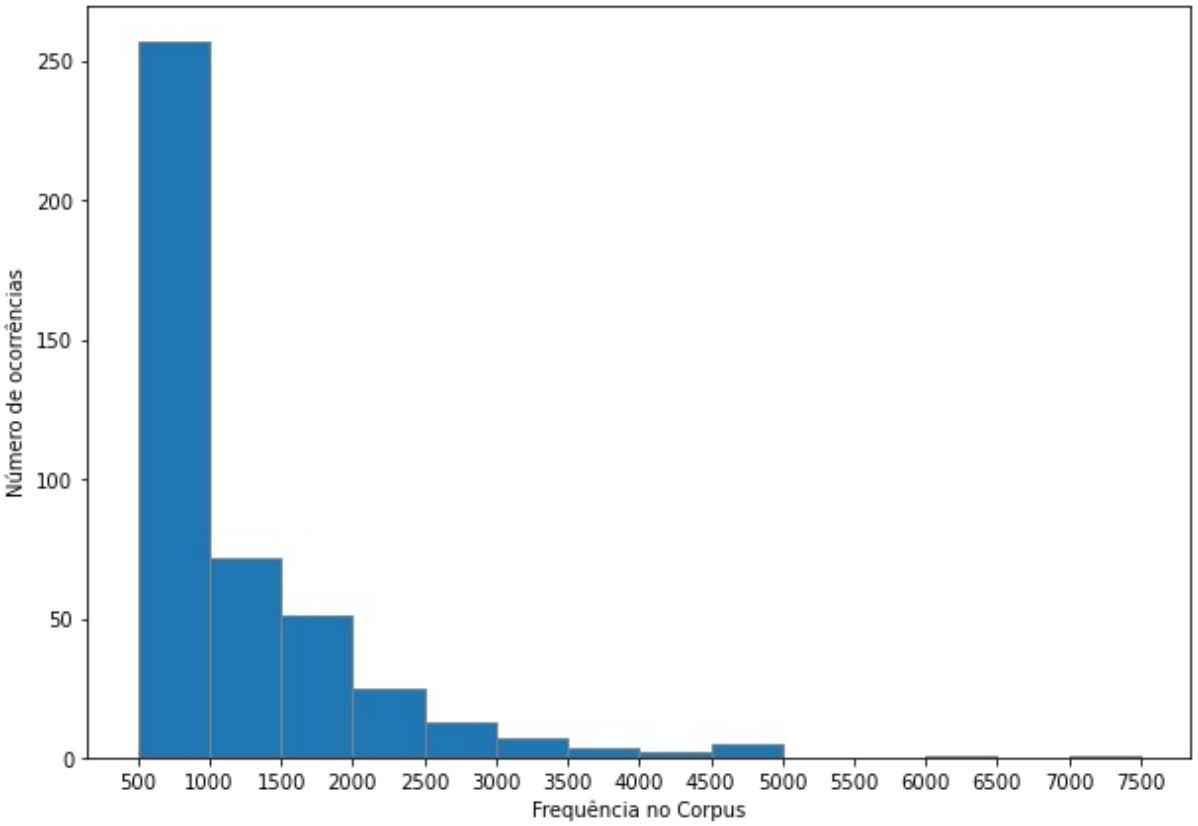
Tabela 1: Palavras mais frequentes do corpus após remoção de *stopwords* e acentuação

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

Na Figura 12 a seguir vê-se a distribuição de frequências para as palavras do corpus. Observa-se que a grande maioria das palavras presentes no corpus possuem baixa frequência, sendo poucas as que ultrapassam 3000 ocorrências ou mais.

Figura 12: Distribuição de frequências das palavras do corpus.



Fonte: Elaborado pelo autor

Por fim foram encontradas as frequências para as palavras selecionadas para a análise, indicadas na Tabela 2.

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.

Para as formas “pai”, “mãe”, “deus”, “homem”,”mulher” e “terra”, foram encontradas palavras de frequência consideradas satisfatórias e são palavras que possivelmente carregam vieses em seus contextos de uso, portanto foram elas as escolhidas para serem analisadas.

3.3 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 a partir de uma modificação da classificação proposta por Bechara (1985).

A delimitação proposta por Bechara (1985) tem início na fase arcaica, que compreende o século XIII até o final do século XIV. Essa fase compreende o período chamado de galego-português, onde os documentos escritos existentes são de variedade culta e erudita. Alguns dos fenômenos encontrados nessa variedade temos:

- possessivos femininos de formas proclíticas (ma, ta, as) ao lado de formas normais (mha, mia; tua, sua), que eram empregados sem muito rigor quanto sua função;
- o -d- etimológico da desinência de 2ª pessoa plural: amades, fazedes, queredes, seeredes, leixedes, fazede, etc.;
- terminação –on (-om) nas formas verbais oriundas de –unt: amáron (amárom), quiseron (quiserom), etc.

A próxima fase é a arcaica média, que corresponde ao intervalo entre a 1ª metade do século XV até a 1ª metade do século XVI. O autor a caracteriza como uma fase de transição, mas destaca a queda do -d- da desinência de 2ª pessoa do plural como essencial para se delimitar esse período (BECHARA, 1985).

A terceira fase proposta por Bechara (1985) é a fase moderna, que vai da 2ª metade do século XVI até o final do século XVII. Alguns dos fenômenos dessa fase são destacados pelo autor:

- a fixação do plural dos nomes em –ão (mãos, cães, leões) e do feminino dos adjetivos em –ão (são / sã);
- a presença obrigatória do pronome demonstrativo antes do pronome relativo em construções como *eu sou o que*, *tu és o que*, *nós somos os que*, etc. (persistindo até final do séc. XVIII).
- a progressiva ação analógica do radical do infinitivo sobre o radical da 1ª pessoa de muitos verbos, como senço / sinto, menço / minto, arco / ardo, etc.

A quarta e última fase definida por Bechara (1985) é a fase contemporânea, que compreende o século XVIII até hoje. Alguns dos fenômenos característicos desse momento são:

- a progressiva eliminação do pronome vós;

- fixação da oposição lhe singular / lhes plural, quando não combinados com os pronomes o, a, os, as;
- o desaparecimento de formas de indeterminação do sujeito como *homem* e *um;*
- o emprego das preposições *per* e *por* é unificado na forma única *por.*

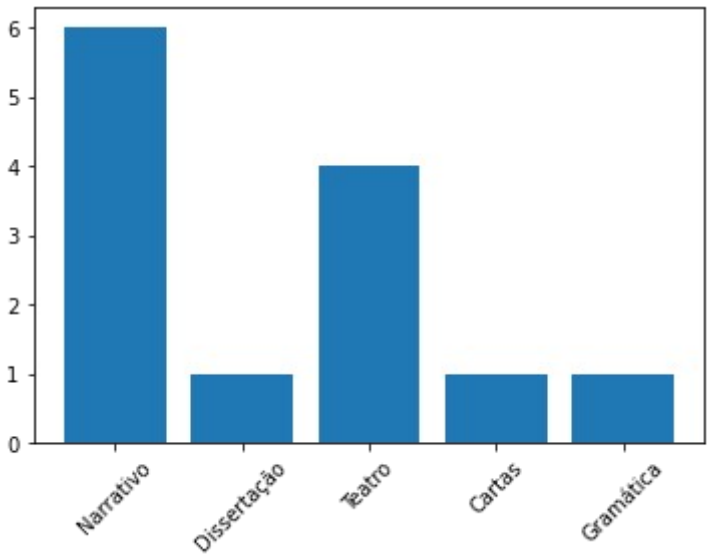
A fase arcaica e a fase arcaica média foram unificadas devido ao baixo número de *tokens* inicialmente obtidos. Foi ainda analisada a quantidade de textos de determinado gênero por período de de tempo analisado, como apresentado nas Figuras 13, 14 e 15.

Tabela 3: Fases do português (adaptado de Bechara(1985)) e respectivo número de tokens no corpus Tycho Brahe.

Fases	**Séculos**	**Nr. de Tokens**
Período I (Arcaica / Arcaica Média)	Até 1ª metade séc. XVI	632.907
Período II (Moderna)	2ª metade séc. XVI até fim séc. XVII	1.230.507
Período III (Contemporânea)	Séc XVII até início séc. XX	1.439.397

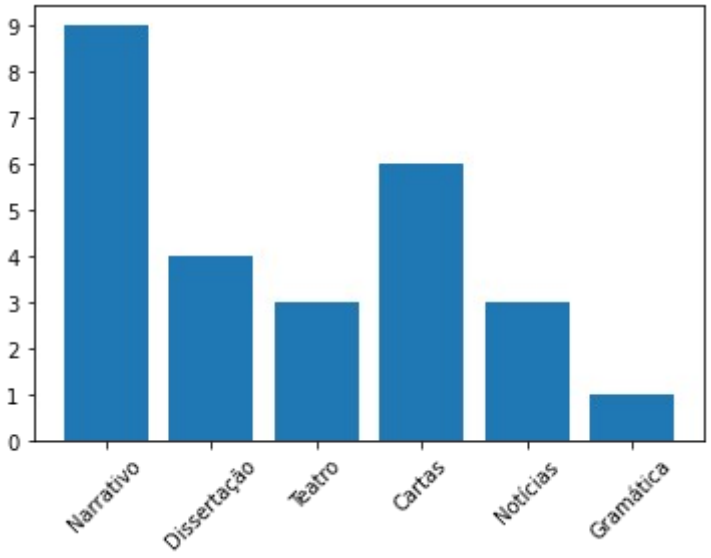
Fonte: Elaborado pelo autor.

Figura 13: Gêneros textuais presentes no período 1



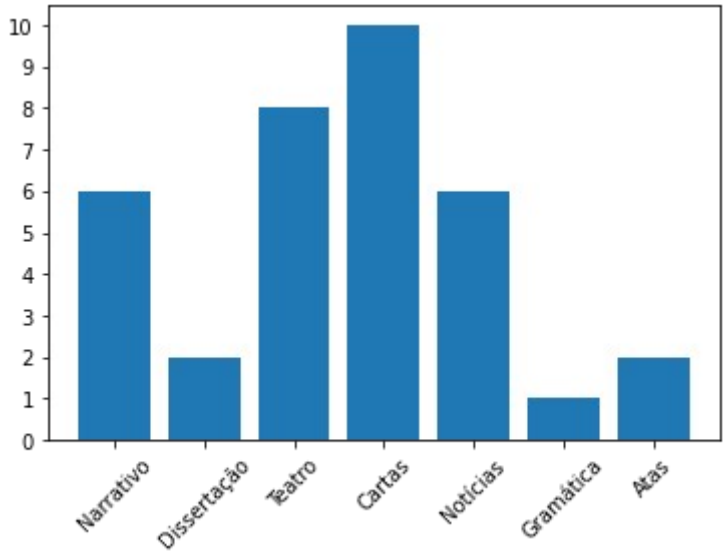
Fonte: Elaborado pelo autor.

Figura 14: Gêneros textuais presentes no período II



Fonte: Elaborado pelo autor.

Figura 15: Gêneros textuais presentes no período III



Fonte: Elaborado pelo autor.

Com um número de tokens melhor balanceado entre os períodos, os textos foram em seguida agrupados de acordo com a data de nascimento do autor e então tratados para o posterior treinamento do modelo.

3.4 LIMPEZA E PROCESSAMENTO DO CORPUS

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

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

A lematização é um processo também comum em tarefas de NLP e consiste em 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. Como exemplo de lematização, tem-se a palavra “professoras”, forma plural e feminina, que, após lematizada, torna-se a palavra “professor”, forma singular e masculina. Esse processo é realizado a fim de manter formas que carregam os mesmos significados agrupadas, assim as formas “andei” e “andou” seriam representadas pelo mesmo *token* “andar”.

Todos os passos foram feitos através da biblioteca Spacy[1] em sua verão 3.1.1.

3.5 TREINAMENTO DOS MODELOS

Os modelos Skip-gram foram treinados, um para cada período, através do código fornecido por Mikolov.[2]

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

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

1 Disponível em spacy.io/

2 Disponível em github.com/tmikolov/word2vec

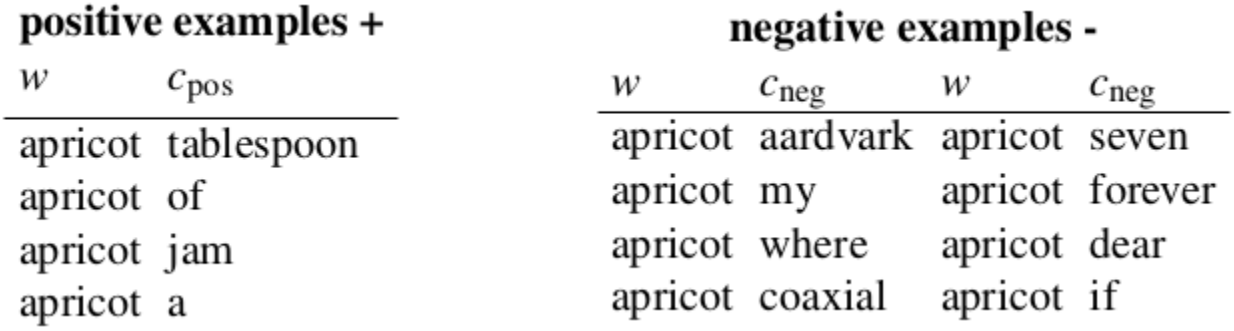
Tabela 4: Hiperparâmetros de treinamento

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. Um exemplo negativo é, neste caso, uma sequência de palavras que não ocorre no corpus. Esses exemplos são gerados aos se substituir uma palavra em exemplo por uma palavra aleatória do corpus. Pode-se ver um exemplo negativo na Figura 16, onde a forma “apricot” é emparelhada em contextos reais na coluna esquerda e em contextos não existentes no corpus na coluna direita. O valor fornecido corresponde à 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.

Figura 16: Exemplos positivos e exemplos negativos para a forma “apricot”



Fonte: Jurafsky; Martin (2008)

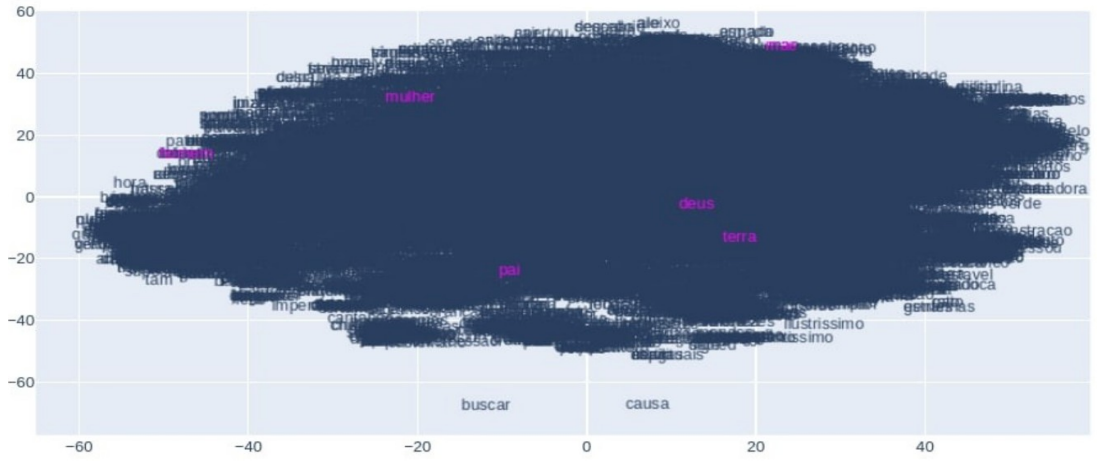
Após a obtenção dos vetores de palavras resta apenas gerar a visualização gráfica.

3.6 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 4.0.1[3].

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 17 a seguir. As imagens finais foram obtidas após a devida ampliação da imagem.

Figura 17: Representação gráfica gerada pelo Gensim



Fonte: Elaborado pelo autor.

3 Disponível em radimrehurek.com/gensim/

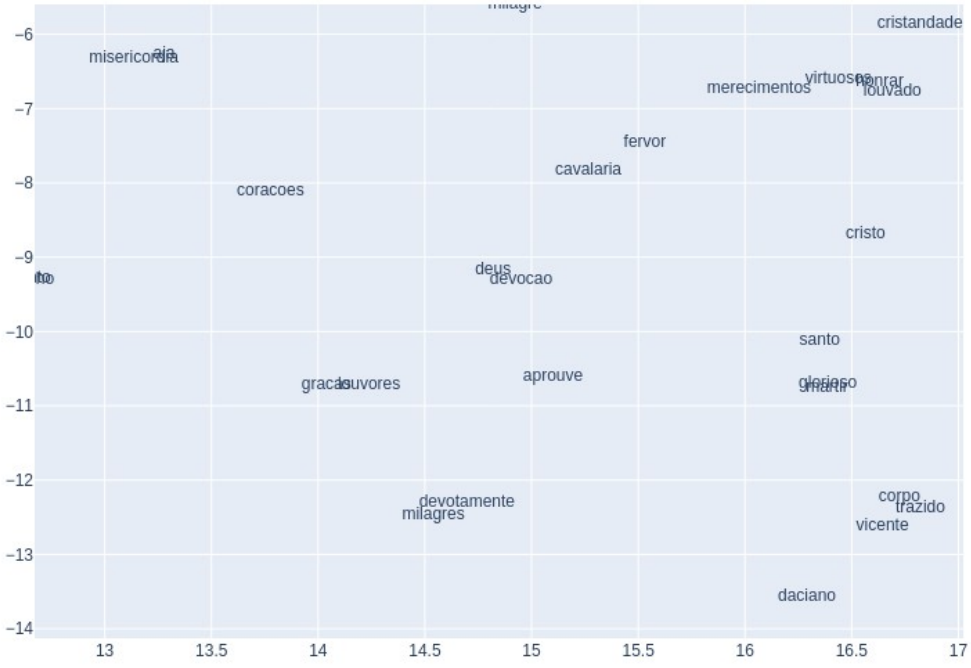
4 RESULTADOS

4.1 REDES DE RELAÇÕES SEMÂNTICAS PARA A PALAVRA “DEUS”

Como primeiro resultado obteve-se as redes de relações semânticas das palavras escolhidas inicialmente para cada um dos períodos delimitados no corpus.

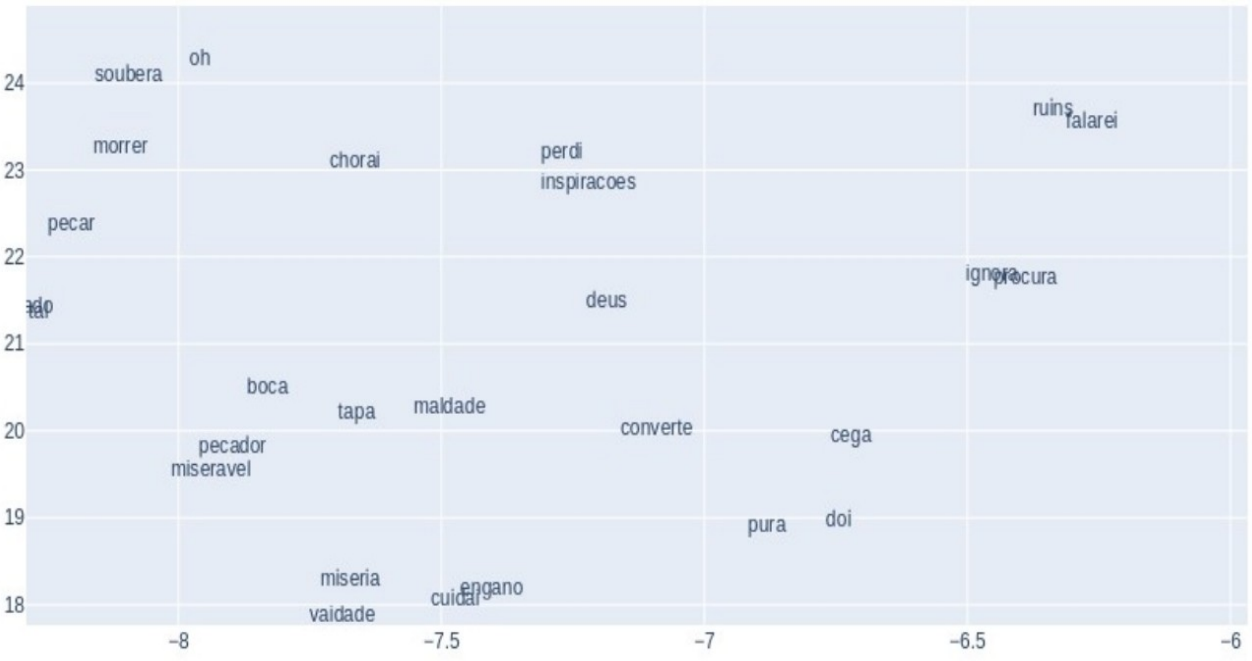
Primeiramente, os gráficos gerados para a palavra “deus” no período I mostram-se adequados ao contexto, com palavras como “devoção”, “louvores”, “devotamente” (Figura 18), em sua proximidade. Já para o período II, vemos palavras como “inspirações”, “converte”, “perdi” e “maldade” (figura 19) próximas. Por fim, no período III têm-se as palavras “amas”, “aflitos”, “misericórdia”, “divinos”, “preconceito”, “santíssima” (figura 20).

Figura 18: Rede de relações semânticas da palavra "deus", período I



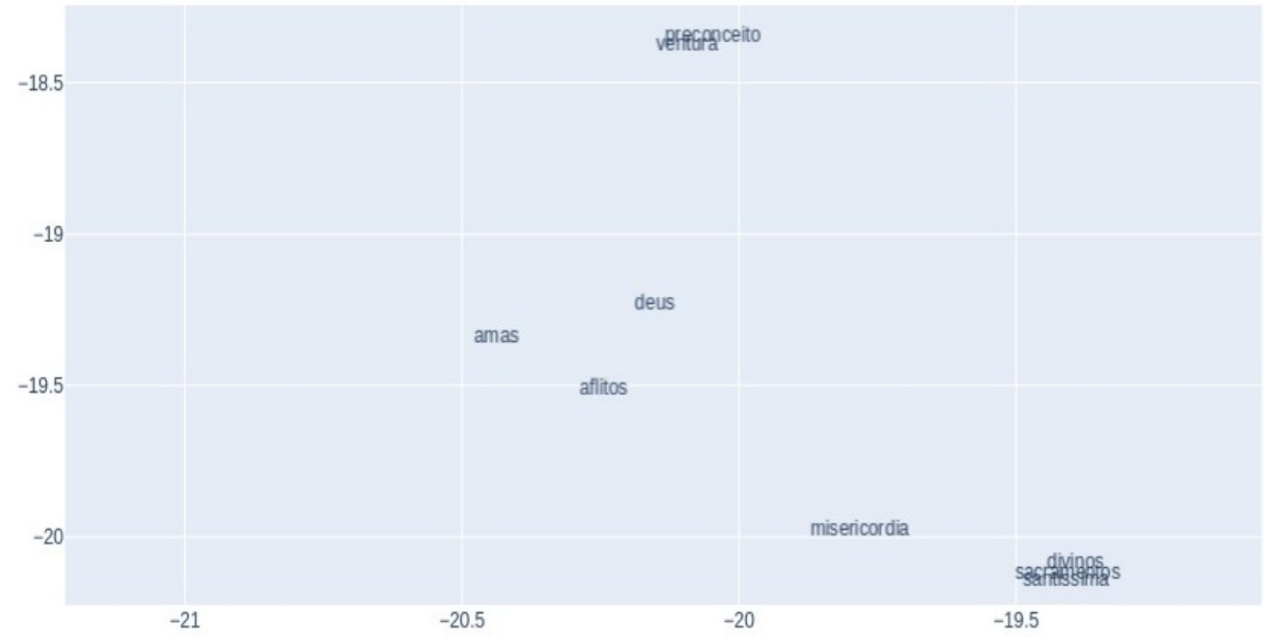
Fonte: Elaborado pelo autor.

Figura 19: Rede de relações semânticas da palavra deus, período II



Fonte: Elaborado pelo autor.

Figura 20: Rede de relações semânticas da palavra "deus", período III

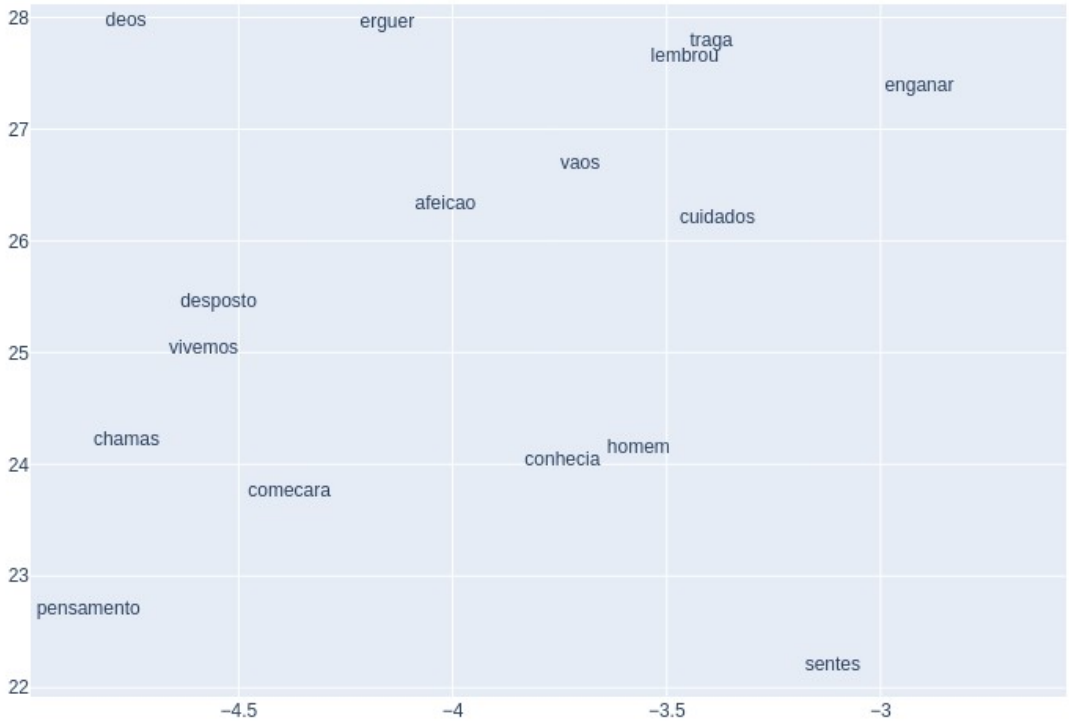


Fonte: Elaborado pelo autor

4.2 REDES DE RELAÇÕES SEMÂNTICAS PARA A PALAVRA “HOMEM”

Para o período I vemos as palavras “conhecia”, “cuidados”, “sentes”, “comecara” mais próximas (Figura 21). Já para o período II têm-se as palavras “conheces”, “falas”, “figueira”, “ouvis”, “ves”, “sejas” (Figura22). No Período III vê-se as palavras “miserável”, “ateu”, “ímpio”, “solene”, “livrar” (Figura 23).

Figura 21: Rede de relações semânticas da palavra "homem", período I

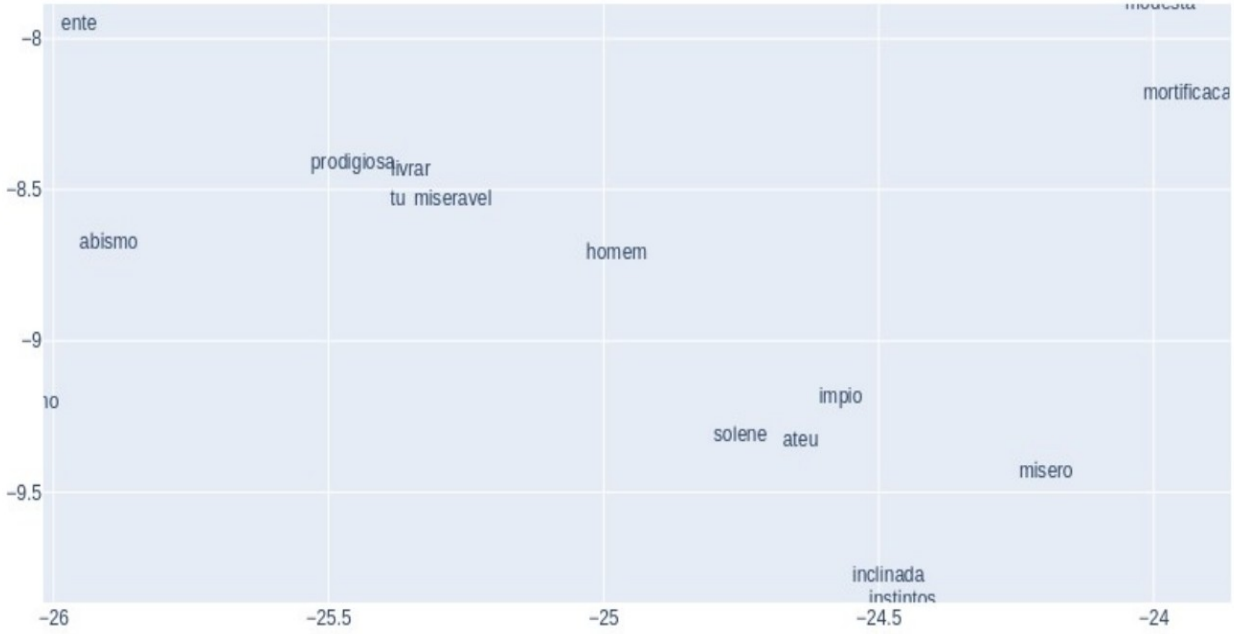


Fonte: Elaborado pelo autor.

Figura 22: Rede de relações semânticas da palavra "homem", período II

Fonte: Elaborado pelo autor.

Figura 23: Rede de relações semânticas da palavra "homem", período III

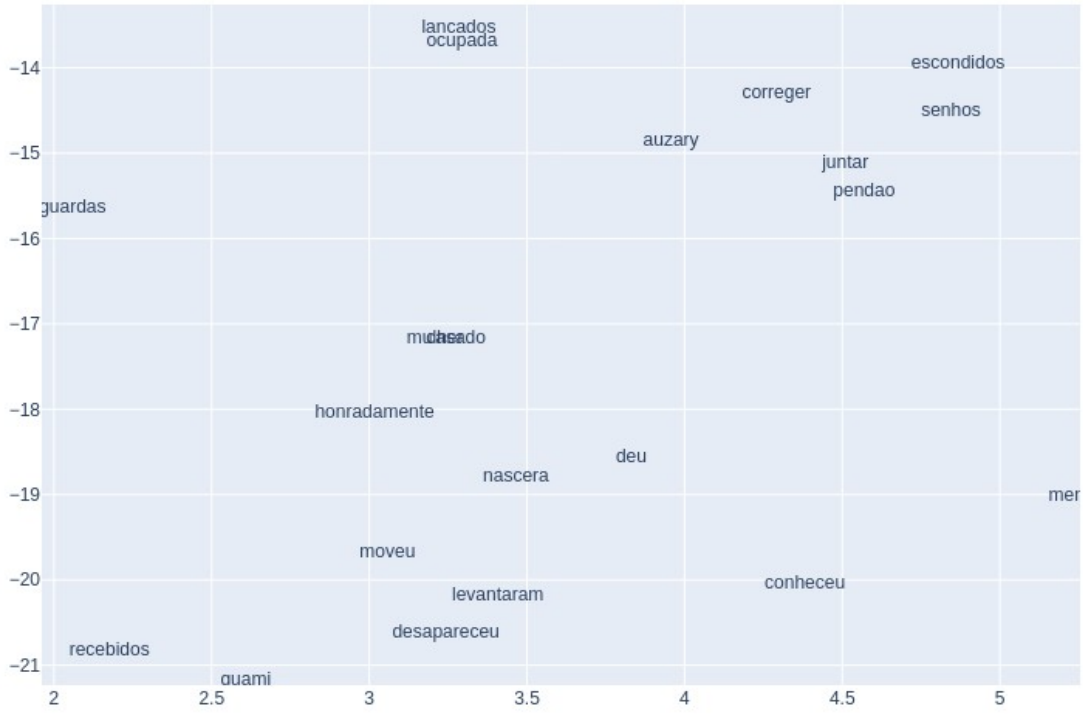


Fonte: Elaborado pelo autor.

4.3 REDES DE RELAÇÕES SEMÂNTICAS PARA A PALAVRA “MULHER”

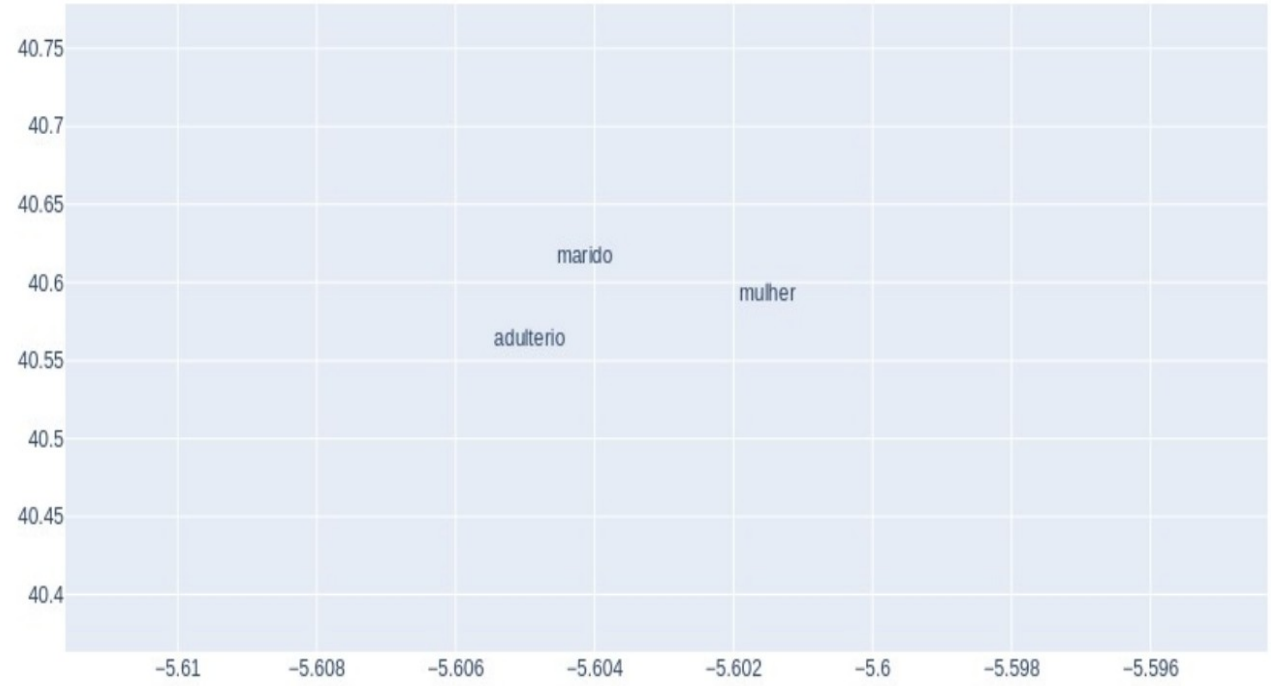
Para a rede da palavra “mulher”, vemos no período I que ela ficou bastante próxima da palavra “casado”, seguida das palavras “nascera”, “deu”, ‘nascera”, “honradamente” (Figura 24). Já no período II, temos as palavras “marido” e “adultério” muito próximas (Figura 25). Finalmente no período III, temos as palavras “desgraçada”, “coitadinha”, “marido” e “margarida” na proximidade da palavra analisada (Figura26).

Figura 24: Rede de relações semânticas da palavra "mulher", período I



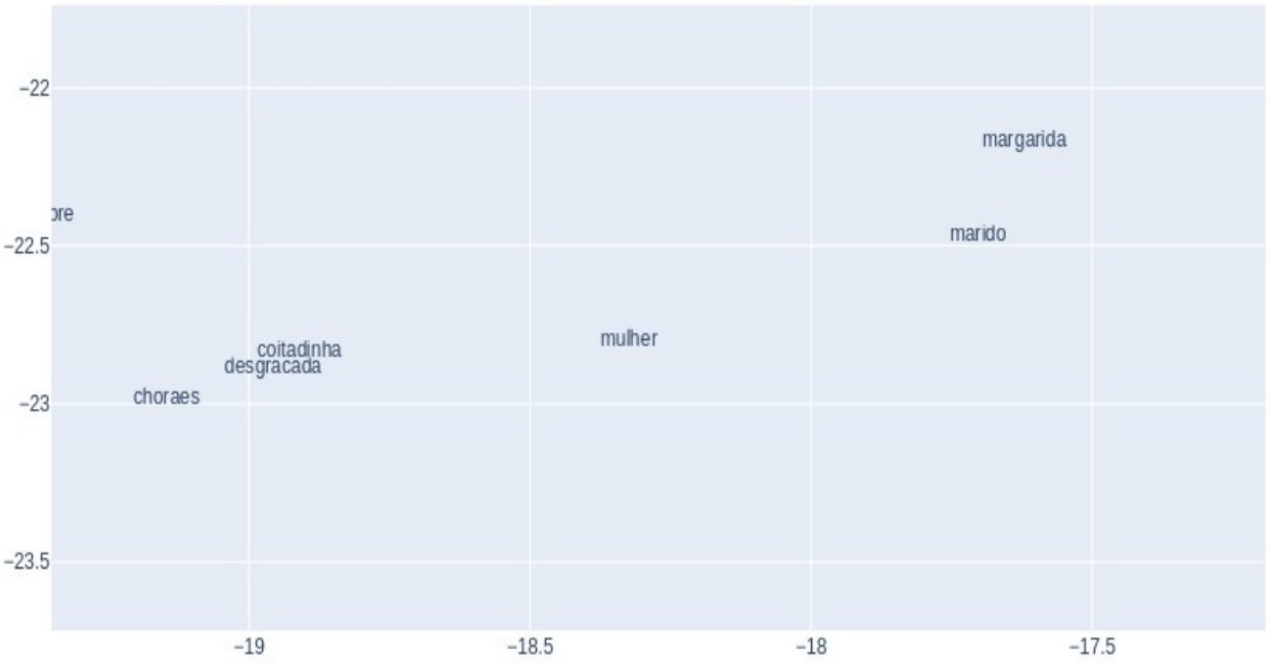
Fonte: Elaborado pelo autor

Figura 25: Rede de relações semânticas da palavra "mulher", período II



Fonte: Elaborado pelo autor.

Figura 26: Rede de relações semânticas da palavra "mulher", período III



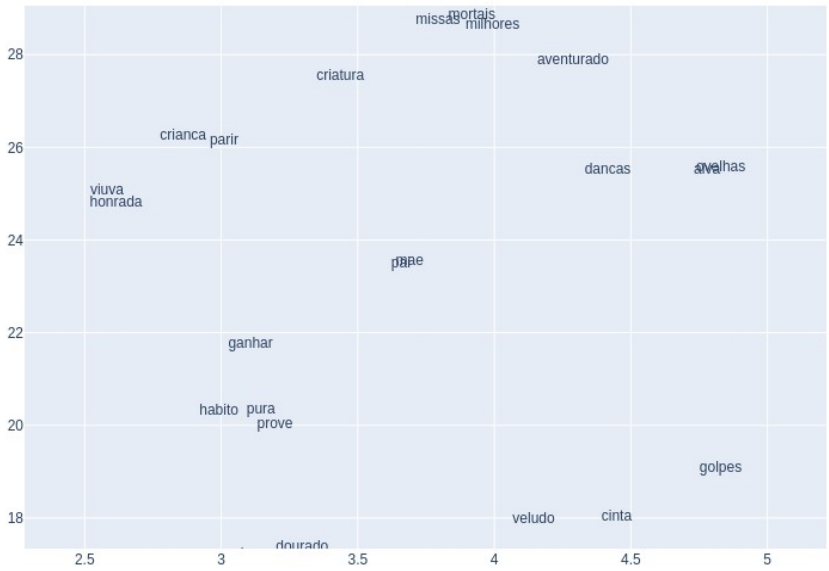
Fonte: Elaborado pelo autor.

4.4 REDES DE RELAÇÕES SEMÂNTICAS PARA AS PALAVRAS “PAI” E “MÃE”

A análise da rede de relações dessas palavras foi realizada em conjunto devido a proximidade que elas se encontram nos resultados. Para a rede no período I, elas se sobrepõe, como mostrado na figura 27. 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 28. Já a palavra “mae” se encontra próxima de expressões como “casados”, “bodas”, “legítimo”, “casal”, “embaracos” e “ajustada”, como visto na Figura 29.

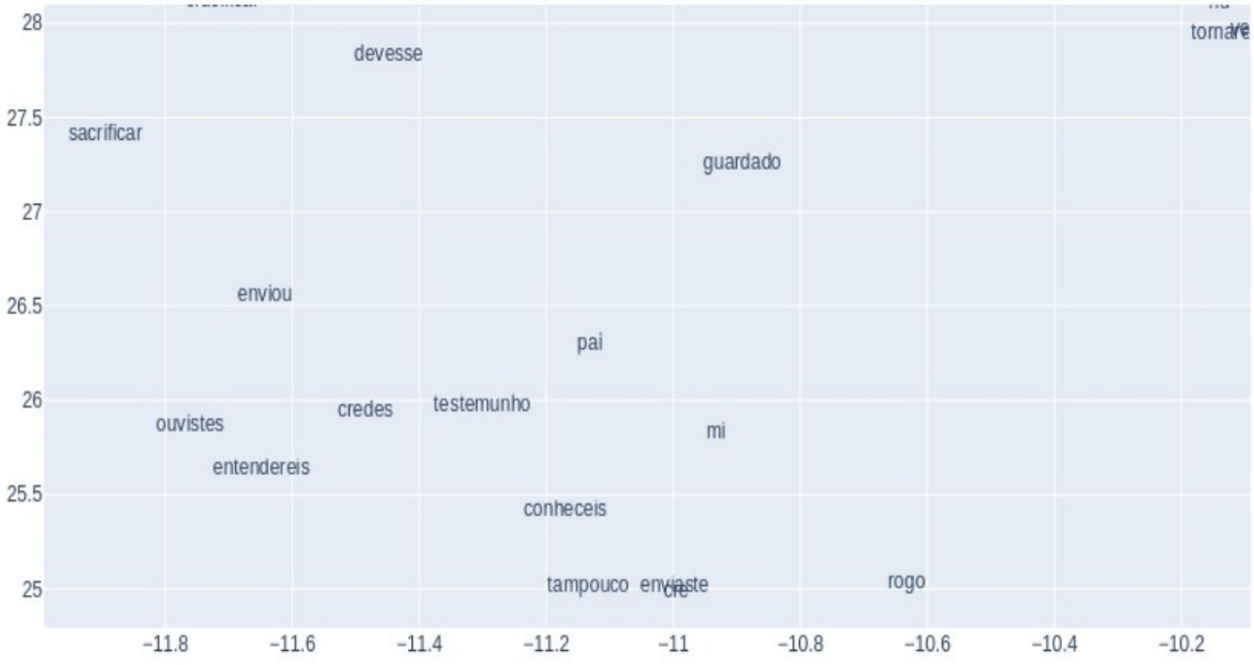
Figura 27: Rede de relações semânticas das palavras "pai" e "mãe", período I.



Fonte: Elaborado pelo autor.

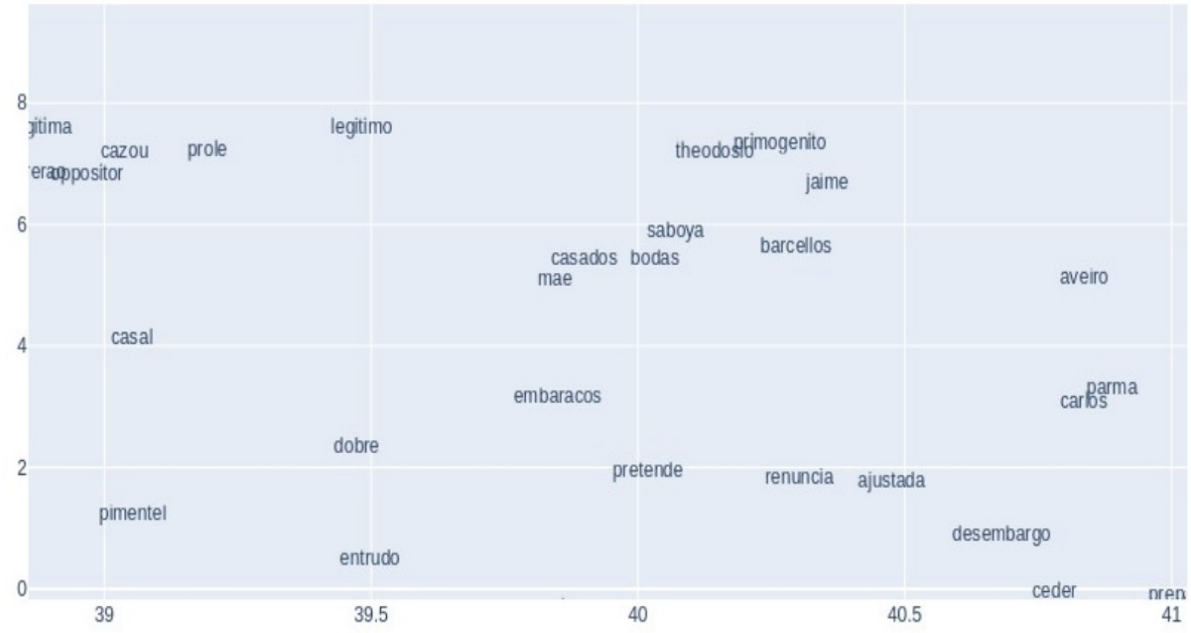
No período III, a rede da palavra “pai” mostra proximidade com “consolar”, “paterno”, “filho”, além do verbo “tourear”, Figura 30. Já o campo de “mãe”, mostra as palavras “virtuosa”, “filha”, “carinhosa”, vide Figura 31.

Figura 28: Rede de relações semânticas da palavra "pai", período II



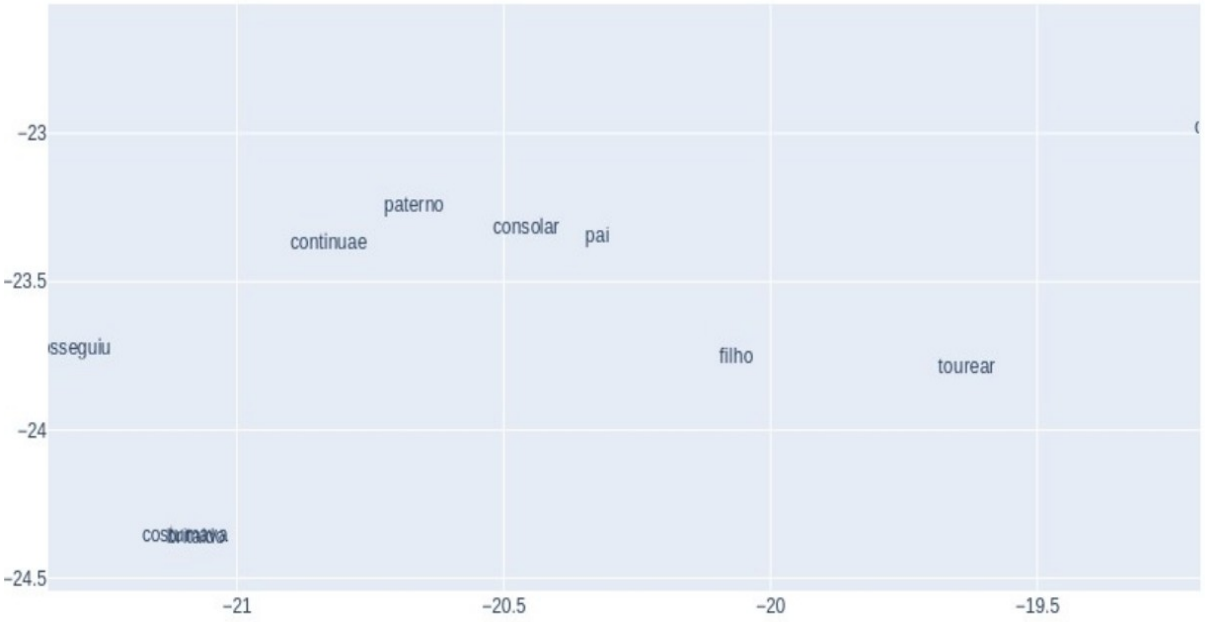
Fonte: Elaborado pelo autor.

Figura 29: Rede de relações semânticas da palavra "mae", período II



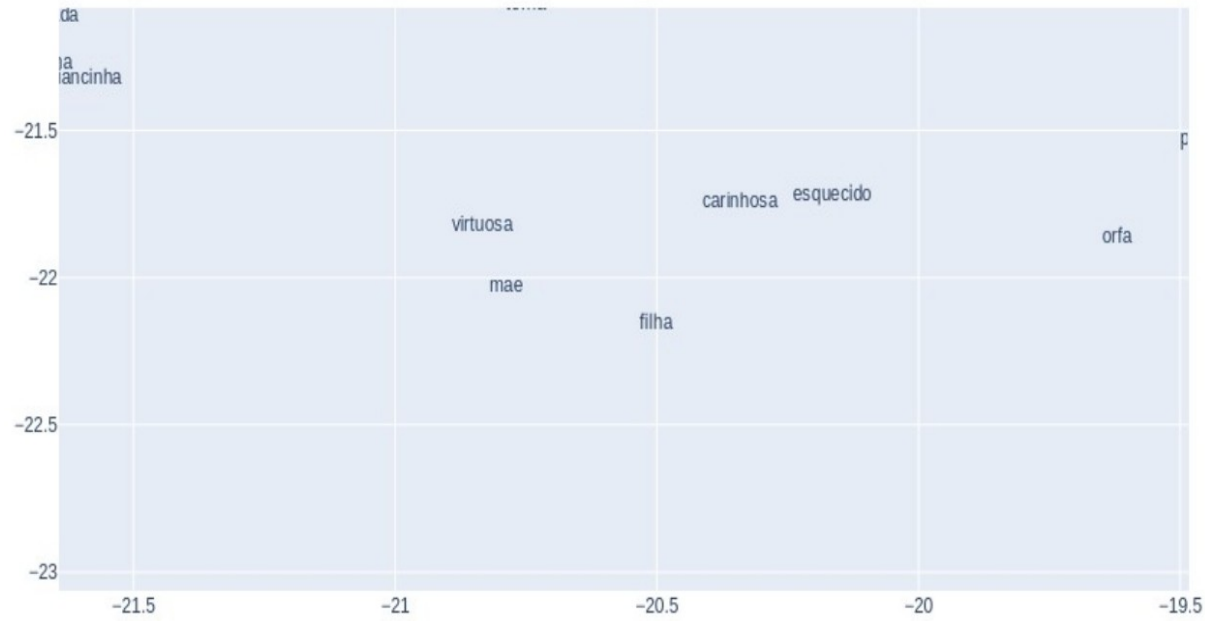
Fonte: Elaborado pelo autor.

Figura 30: Rede de relações semânticas da palavra "pai", período III



Fonte: Elaborado pelo autor.

Figura 31: Rede de relações semânticas da palavra "mae", período III

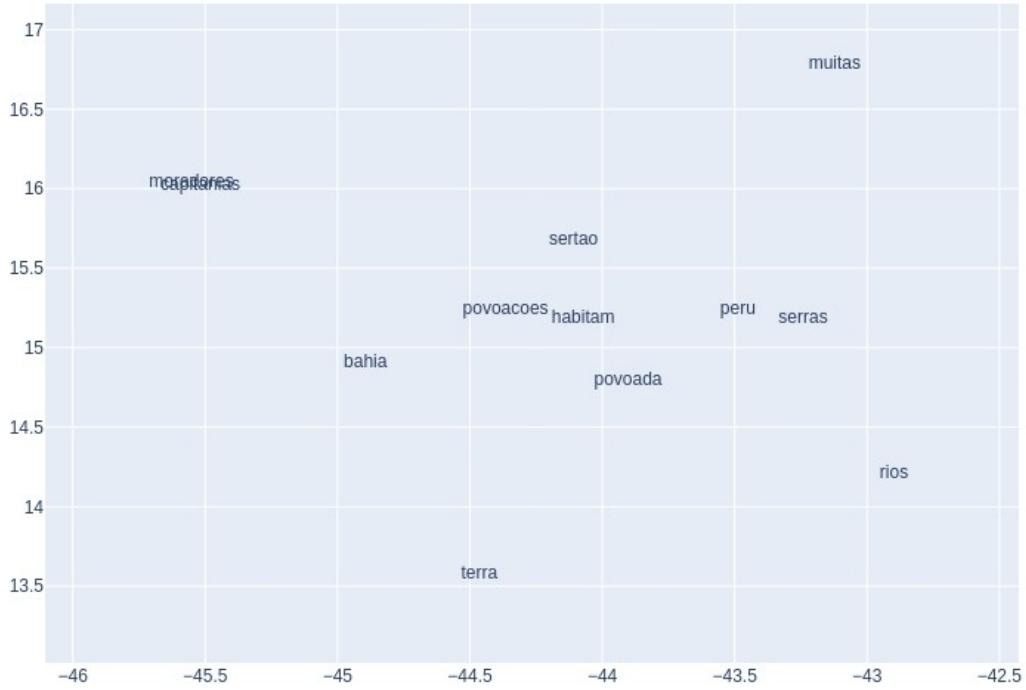


Fonte: Elaborado pelo autor.

4.5 REDES DE RELAÇÕES SEMÂNTICAS PARA A PALAVRA “TERRA”

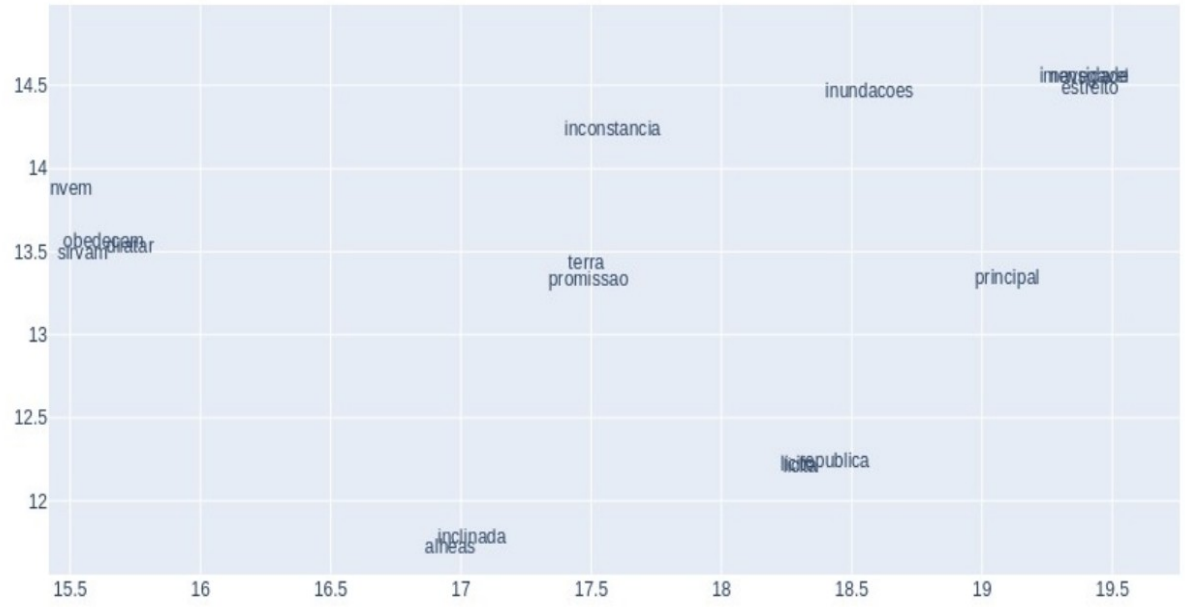
Para o período I, a palavra “terra” possui em sua vizinhança as palavras “sertão”, “habitam”, “povoada”, “bahia”, “rios”, conforme indicado na Figura 32. Já para o período II a figura 33 mostra as palavras “inconstância”, “promissão” e destaca-se a palavra “republica”. Finalmente o período III apresenta palavras como “campinas”, “montanhas”, “ribeiras”, “ventos” e “tempestades, como visto na figura 34.

Figura 32: Rede de relações semânticas de "terra", período I



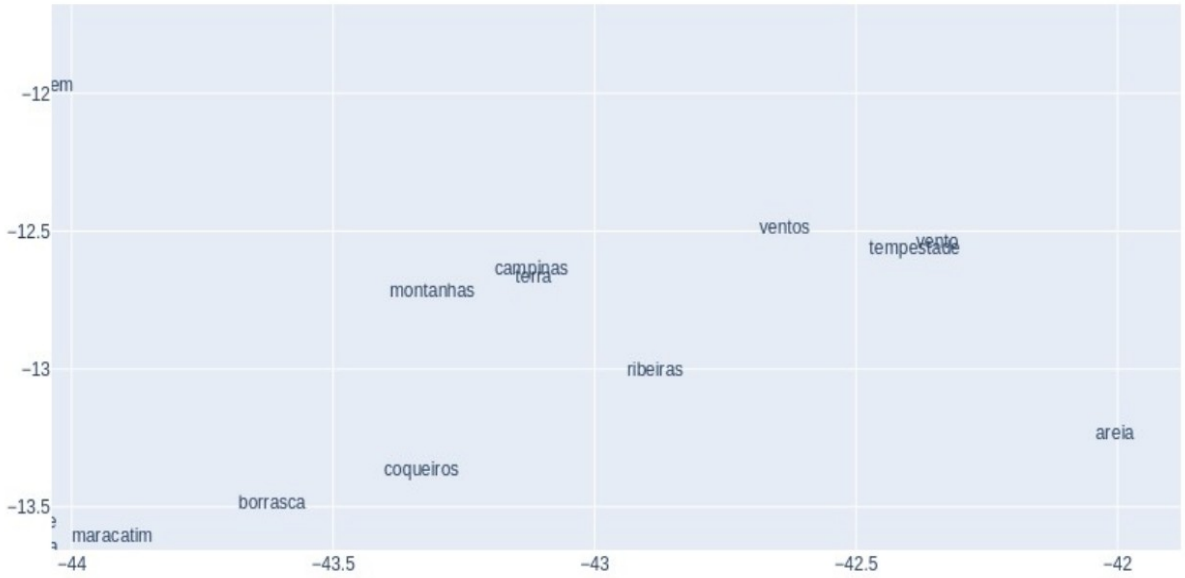
Fonte: Elaborado pelo autor

Figura 33: Rede de relações semânticas de "terra", período II



Fonte: Elaborado pelo autor

Figura 34: Rede de relações semânticas para "terra", período III



Fonte: Elaborado pelo autor

4.6 DISCUSSÃO DOS RESULTADOS

Inicialmente vê-se que as redes de relações passaram por mudanças de acordo com os períodos analisados. Observa-se para a palavra “deus” que ela surge em contextos ligados a religião, o que é um resultado esperado.

A palavra “homem” não vem acompanhada, em um primeiro momento, de palavras que aparentam estar fortemente ligadas a ela. Já no segundo período vemos a palavra associada a verbos sensoriais como “vedes”, “ouvis”, “falas”. Essa ausência de palavras fortemente ligadas a forma “homem” pode ser uma consequência de seu uso como forma de indeterminação do sujeito nesses períodos, como informado por Bechara (1955). Já no último momento vê-se a forma “homem” em palavras de viés negativo como “miserável”, “impio” e também a palavra “ateu” e a palavra com viés positivo “solene” (BECHARA, 1985).

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 à palavra “mulher” mostrando a relação com o “gênero”.

Para as palavras “pai” e “mãe aparecem bastante próximas em sua rede de relações. 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 contextos geográficos, como “bahia” e “sertão”. Já no período II a palavra parece ser encontrada em contextos diferentes, tendo em vista o surgimento da palavra “república”. E por fim

ela se encontra novamente relacionada a termos geográficos como “campinas”, “montanhas”, “ventos”, “areia’.

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 sua rede. 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, o processo de lematização não foi eficiente. Encontram-se formas verbais flexionadas, como “vedes”, “sejas”, “sabes”, “amas” e formas adjetivais apresentando o gênero feminino como em “coitadinha”, “desgraçada”, enquanto eram esperadas suas formas dicionarizadas. A lematizador utilizado é fornecido pela biblioteca Spacy, que realiza o processo automaticamente e possui acurácia relativamente baixa (76%). Além disso, a lematização do corpus diacrônico sofreu também por possuir formas desconhecidas ao modelo, como a palavra “molher”, e o verbo “cazar”, que utiliza um conjunto de regras para gerar os lemas.

5 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 da técnica de NLP conhecida como w*ord embeddings* (vetorização de palavras), que permite 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 mais pobre.

Os obstáculos para um melhor resultado se dão principalmente pelo tamanho do corpus. Os resultados apresentados por Hamilton, Leskovec e Jurafsky (2016) foram obtidos através de um corpus que possui mais de 410 milhões de tokens, quando separados por períodos são mais de 100 milhões de tokens, um valor muito distante da quantidade de tokens obtidos com o corpus Tycho Brahe. Apesar de os autores citarem outras formas de obtenção de w*ord 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, e além disso esses diferentes gêneros encontram-se desbalanceados quanto a sua representação no corpus.

Considerando as limitações encontradas nesta análise, propõe-se para estudos futuros o uso de formas alternativas de obtenção de w*ord 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, até o momento e até onde o autor constatou, uma análise quantitativa que utilize metodologia similar para dados diacrônicos.

6 REFERÊNCIAS

BĂLAN, O. et al. Emotion classification based on biophysical signals and machine learning techniques. **Symmetry**, v. 12, n. 1, p. 21, 2020.

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

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit**. [s.l.] O'Reilly Media, Inc., 2009.

BOCHKAREV, V.; SOLOVYEV, V.; WICHMANN, S. Universals versus historical contingencies in lexical evolution. **Journal of The Royal Society Interface**, v. 11, n. 101, p. 20140841, 2014.

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.

FIRTH, J. R. A synopsis of linguistic theory, 1930-1955. **Studies in linguistic analysis**, 1957.

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. **st International Conference on Learning Representations**, 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.

LABOV, W. Some principles of linguistic methodology. **Language in society**, v. 1, n. 1, p. 97–120, 1972.

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

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **1st International Conference on Learning Representations**, ICLR 2013.

OSGOOD, C. E.; SUCI, G. J.; TANNENBAUM, P. H. **The measurement of meaning**. [s.l.] University of Illinois press, 1957.

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.

YAO, Z. et al. **Dynamic word embeddings for evolving semantic discovery**. . In: PROCEEDINGS OF THE ELEVENTH ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING. 2018.