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

## *Semantic change and word embeddings: case studies on the diachrony of Portuguese*

Lucas Fonseca Lage
Universität des Saarlandes (UdS), Saarbrücken, Saarland / Alemanha
flage.lucas@gmail.com
https://orcid.org/0000-0003-1141-4236

Evandro Landulfo Teixeira Paradela Cunha
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais / Brasil
cunhae@ufmg.br
https://orcid.org/0000-0002-5302-2946

**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 investigar a viabilidade de tais métodos para análise de mudança semântica. Para isso, foram analisadas ocorrências das formas "deus", "homem", "mulher", "pai", "mae" e "terra" no *corpus* Tycho Brahe do português. Através do algoritmo Skip-gram foram gerados Word Embeddings, e, posteriormente, visualizações para a rede de relações semânticas de cada palavra em três diferentes recortes temporais. Através das visualizações foram observadas evidências da organização semântica do léxico, além de sua reorganização através do tempo.

eISSN: 2237-2083

**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 concepts which are relatively stable in time, socially shared and well encoded. They are well organized in a network where similar concepts are grouped next to each other. On a similar note, the lexicographer Georges Matoré proposes associative relationships between words and defines the concepts of notional field and testimonial words, which are organizational elements of the lexicon. Using computational techniques such as Word Embeddings, which represent words as vectors in a vector space, it is possible to analyze groupings of words based on their semantic features. This paper aims to explore the viability of such methods in semantic change. The occurrences of the word forms "deus", "homem", "mulher", "pai", "mae" and "terra" were analyzed in the Tycho Brahe *corpus* for Portuguese. Word Embeddings were created using the Skip-gram algorithm, and visualizations for a semantic feature network were created for each word in three different time slices. Evidence of the semantic organization of the lexicon and its reorganization was observed through the generated visualizations.

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

Recebido em 27 de fevereiro de 2022

Aceito em 31 de maio de 2022

## 1 Introdução

Muitas são as propostas a respeito do que causa a mudança de significado de itens lexicais. Uma delas é a de Givón (1995), que caracteriza o léxico como um repositório de conceitos relativamente estáveis no tempo, compartilhados socialmente e bem codificados. De acordo com sua perspectiva, esses conceitos são interconectados em rede, de forma que a ativação de um conceito leva à ativação de conceitos vizinhos.A linguagem teria evoluído em paralelo com mecanismos cognitivos, com a organização sociocultural e com as habilidades comunicativas dos hominídeos. Assim, em uma sociedade na qual 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).

Antes mesmo de Givón propor suas ideias e, com um olhar voltado para a lexicografia, Georges Matoré apresentou em seu trabalho 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 certos termos e busca, também, analisar como essas redes se alteram temporalmente. Entretanto, devido às fortes críticas à sua metodologia, os estudos em lexicologia social foram negligenciados, tendo sido retomados, mais recentemente e com uma abordagem renovada (a lexicologia sócio-histórica), por Cambraia e colaboradores (CAMBRAIA, 2013; DORES; TOLEDO, 2018; RAFAEL; SIMIÃO, 2019; entre outros).

Tendo em vista os recentes desenvolvimentos de técnicas na área da computação, em especial as técnicas de *word embeddings* (vetorização de palavras), novas categorias de análise linguística têm se tornado possíveis. Por meio de estudos na área de linguística de *corpus*, que têm disponibilizado dados de qualidade em abundância, tornam-se viáveis novas metodologias de pesquisa que permitem analisar uma grande quantidade de dados (BERBER SARDINHA, 2004).

As técnicas de manipulação de dados muitas vezes não precisam ser sofisticadas, como se observa no trabalho de Michel *et al*. (2011), no qual, por meio de medidas de frequência e análise de entidades culturalmente relevantes, são expostos fenômenos sobre a evolução da gramática, relevância cultural e até mesmo censura. Alguns exemplos são a constante competição entre formas regulares e irregulares do passado no inglês, a redução no tempo em que entidades famosas permanecem relevantes, e a supressão de nomes judeus em livros alemães durante a década de 1930. Essas mesmas medidas podem auxiliar tanto na decisão de inclusão de novos termos em um dicionário quanto na remoção de termos irrelevantes (MICHEL *et al*., 2011).

A frequência de uso de palavras também pode ser vista como evidência para fenômenos de mudança. Givón 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) sugerem que a mudança lexical é favorecida não só por mudanças em um ambiente social e natural, como também em um 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 as duas variedades linguísticas se tornaram mais próximas dado ao advento da mídia de massa, que cresceu exponencialmente nos séculos XX e XXI (BOCHKAREV; SOLOVYEV; WICHMANN, 2014; GIVÓN, 2001).

Usando técnicas computacionais mais complexas, Hamilton, Leskovec e Jurafsky, (2016) utilizam vetorização de palavras não só para buscar formas que sofreram mudança de significado, mas também para validar mudanças de significado já conhecidas. Com sua análise, é possível, ainda, buscar as formas que passaram por maiores mudanças semânticas. Após o uso dessas técnicas e com a validação de seus próprios trabalhos, os autores chegam a 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. Portanto, o algoritmo desenvolvido é capaz de codificar associações como "apple" e "strawberry" em um primeiro ponto no tempo, e em outro ponto, "apple" e "iphone" (YAO *et al*., 2018).

O presente artigo visa avaliar as mudanças semânticas sofridas por expressões ou palavras por meio de técnicas desenvolvidas na área de Processamento de Língua/Linguagem Natural (PLN). A princípio, utiliza-se a 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 nas quais 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 vetorização de palavras foram fundamentados, indiretamente, por autores como Georges Matoré e Talmy Givón (CAMBRAIA, 2013; GIVÓN, 1995).

Como fonte de dados, é utilizado o Corpus Histórico do Português Tycho Brahe, que abrange textos do século XIII até o século XX (DE SOUSA, 2014; GALVES; ANDRADE; FARIA, 2017).

## 2 O léxico em Givón e o léxico em rede

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), as abordagens funcionalistas tratam a estrutura gramatical da língua em relação aos 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 em que 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, de acordo com essas abordagens, os dados analisados devem ser obtidos a partir de produções reais de fala e escrita (CUNHA, 2008).

Na perspectiva funcionalista, destacam-se os estudos de Givón (1995, 2001), que desenvolveu uma gramática propriamente funcionalista. Em sua obra, ele 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 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 (GIVÓN, 1995). São eles o Sistema de Representação Cognitiva e os Sistemas de Codificação.

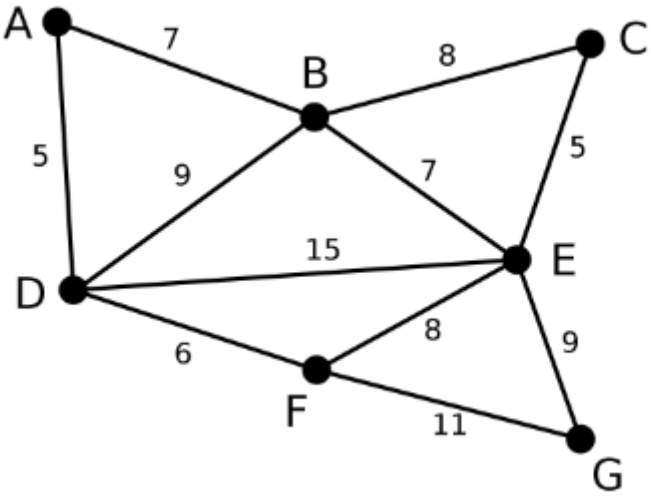
Dentro do Sistema de Representação Cognitiva, tem-se três componentes: o léxico conceitual, a informação proposicional e o discurso multi-proposicional. Aqui, interessa principalmente a definição de léxico conceitual, pois é a partir dela que Givón relaciona, inicialmente, o léxico ao meio e às experiências humanas (GIVÓN, 1995).

O léxico humano é definido na mesma obra 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 universo 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 Nessa visão, ser estável no tempo significa que as palavras e os conceitos associados a elas não estão em um fluxo rápido – por exemplo, 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 (2001) conclui que ele está organizado por meio 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 de como se organiza a rede descrita por Givón, que se assemelha a um grafo. Grafos são uma abstração matemática para se representar objetos e as relações entre eles, os quais possuem nós conectados por arestas. As arestas de um grafo podem possuir pesos, o que pode corresponder à distância entre dois pontos, por exemplo. Na Figura 1, observam-se os nós de "A" a "G" e as arestas com diferentes pesos conectando-os. 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; GRIFFIN, 2017; SWINNEY, 1979).

Figura 1 - Exemplo de grafo. Cada letra do alfabeto corresponde a um nó, enquanto os números indicam os pesos das arestas




Fonte: Griffin (2017, p. 47).

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 interconectados de nós. 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 corresponderia, então, a um substantivo. Pode se referir, ainda, 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 autores já haviam buscado trabalhar o léxico de forma sistemática. Um deles é Georges Matoré, que trabalhou com o léxico como uma rede, buscando uma lexicologia social. De acordo com Cambraia, através de seu livro *La lexicologie sociale*,

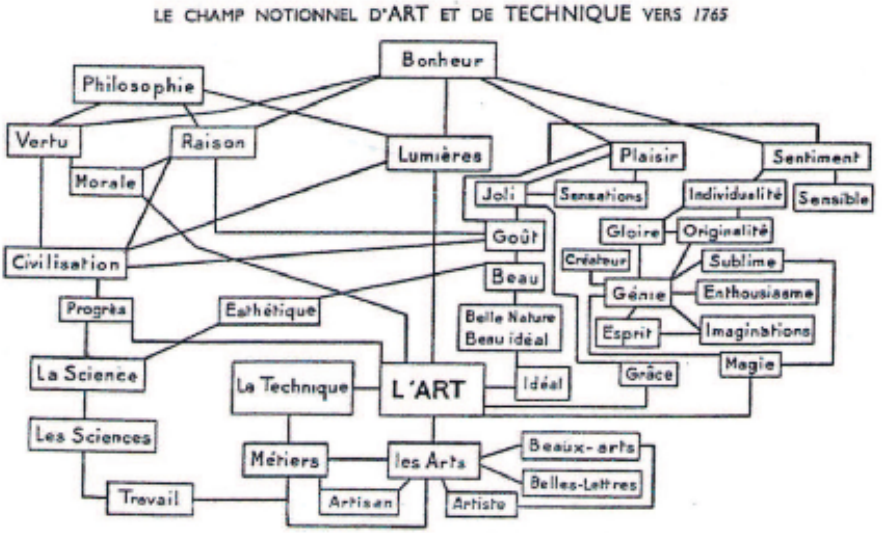
Georges Matoré (1949) cita a influência de fatores sociais no estudo do léxico. Ele propõe uma série de princípios para uma nova lexicologia, denominada lexicologia social. O primeiro desses princípios propõe que forma e conceito são indissociáveis. Segundo o autor, a formação de uma palavra equivale 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).

Matoré mantém uma visão sistêmica do comportamento do vocabulário, 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, 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 2, e o campo nocional de Artista entre os anos de 1827 e 1834, Figura 3 (CAMBRAIA, 2013).

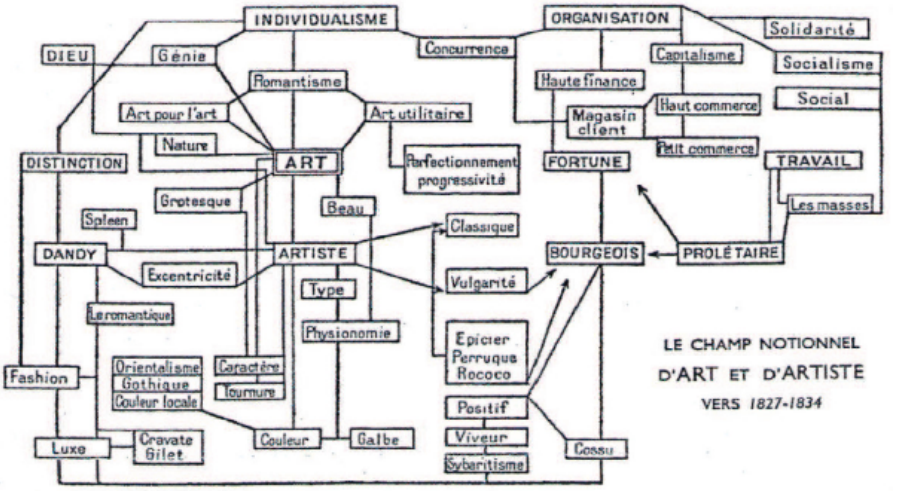
Figura 2 - Campo nocional de “Arte” e “Técnica” em 1765, segundo Georges Matoré




Fonte: Cambraia (2013, p. 163).

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 dos termos “palavras-testemunho” e “campos nocionais”. Além disso, uma consideração muito importante foi realizada por Robin. A autora considera que os estudos de Matoré não poderiam refletir a sociedade como um todo, mas apenas os grupos aos quais 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 (CAMBRAIA, 2013; ROBIN; DE MENESES BOLLE, 1977).

Figura 3 - Campo nocional de "Artista" entre 1827-1834, segundo Georges Matoré (2013)




Fonte: Cambraia (2013, p. 164).

Cambraia (2013) apresenta um estudo de caso, levando em consideração elementos metodológicos que devem constituir uma lexicologia sócio-histórica. Entre esses elementos tem-se a articulação entre fatores intralinguísticos e extralinguísticos, que, até aquele 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.

## 3 Semântica vetorial e *word embeddings*

O conceito de *word embeddings*, ou vetorização de palavras, assim como as propostas de Givón, também se desenvolveu 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 como John R. Firth, na década de 1950, quando se percebeu que palavras sinônimas ocorrem no mesmo ambiente/contexto. Firth afirma que "Você conhece uma palavra pela sua companhia"[1] (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 entre os contextos onde ocorrem. Ambas são substantivos, referem-se a animais domésticos, de quatro patas, alimentados pelos seus donos, mas um mia e outro late. Essas semelhanças semânticas têm como consequência seu surgimento em contextos sintáticos semelhantes. A similaridade entre palavras, sentenças ou documentos é bastante útil em diversas tarefas de PLN como sistemas de resposta a perguntas, paráfrase e sumarização (JURAFSKY; MARTIN, 2008).

Além de relações entre palavras como sinonímia, polissemia, antonímia e similaridade, as 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, no qual 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 observam-se as valorações das palavras *courageous*, *music*, *heartbreak* e *cub* em três eixos. O eixo de valência está relacionado à agradabilidade do estímulo gerado, o de excitação diz respeito à intensidade da emoção provocada pelo estímulo e o eixo de dominância se refere ao grau de controle exercido pelo estímulo (JURAFSKY; MARTIN, 2008; OSGOOD; SUCI; TANNENBAUM, 1957).

---

[1] "You shall know a word by the company it keeps" (FIRTH, 1957, p.11).

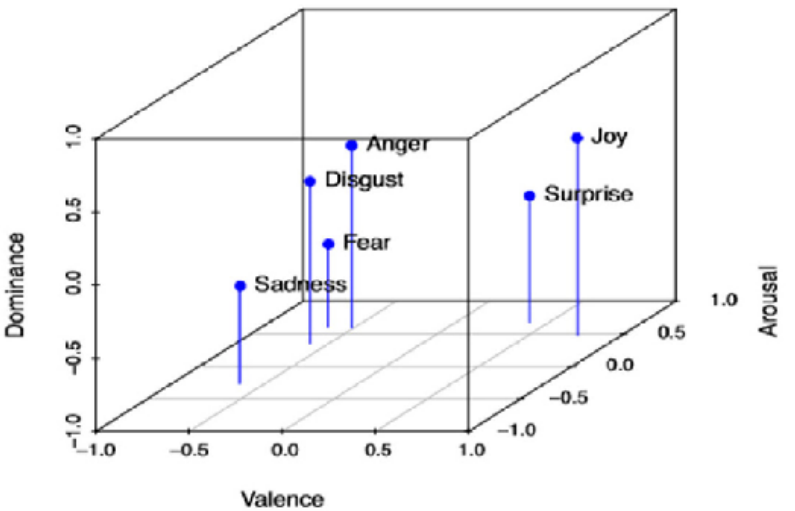
Figura 4 - Variação quantitativa do sentido afetivo em três eixos

| | 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, p. 106)

A grande contribuição de Osgood para o campo foi a percepção de que as palavras poderiam ser representadas em um espaço vetorial, 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 (nesse caso, de três dimensões). As palavras mostradas são *anger, disgust, fear, joy, sadness* e *surprise* ("raiva", "repulsa", "medo", "alegria", "tristeza" e "surpresa").

Figura 5 - Representação afetiva de palavras, segundo Osgood et al. (1957)




Fonte: Bălan *et al*., (2020, p. 4).

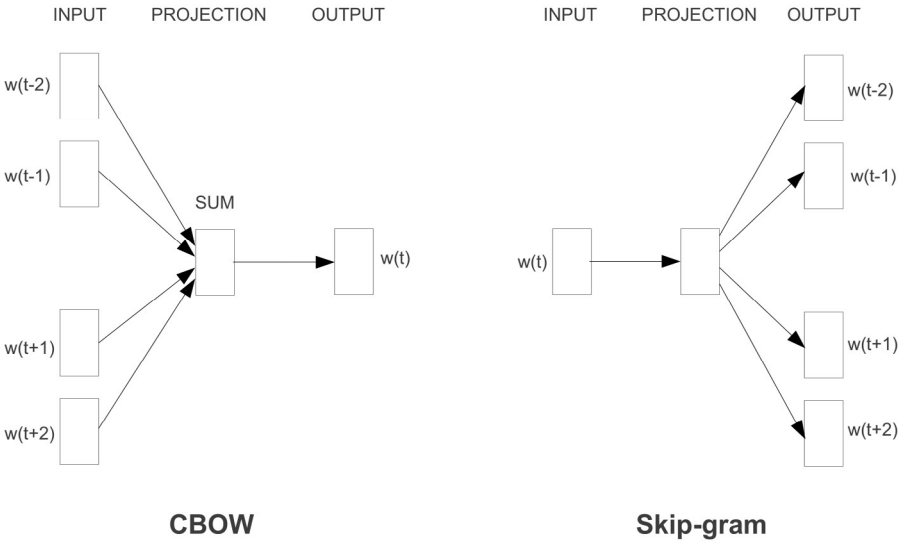
Após esses estudos iniciais, os métodos para obtenção de representações vetoriais evoluíram até chegarem a usar métodos de aprendizado automático, como os modelos *Continuous Bag-of-Words* e *Skip-gram*.

Um dos maiores avanços em vetorização de palavras é a publicação de Mikolov *et al*. (2013). Nela, é proposta uma arquitetura de redes neurais capaz de criar uma representação vetorial para cada palavra apresentada ao modelo, e, em seguida, é possível reproduzi-las em um espaço vetorial. O mérito desse trabalho está relacionado à 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)*. A primeira é criada a partir de uma tarefa de predição, em que uma palavra é prevista dado seu contexto, ou palavras vizinhas, como entrada em uma rede neural. O contexto, no caso, deve ser entendido como as palavras em posições anteriores e posteriores a 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 esteja 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. Na Figura 6 tem-se uma representação da arquitetura dos modelos aqui discutidos (MIKOLOV *et al*., 2013).

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




Fonte: Mikolov *et al.* (2013, p. 5)

Para a validação dos modelos criados, os autores desenvolveram tarefas de predição baseadas em relações sintáticas e semânticas. Esses testes são usados como uma métrica de desempenho dos modelos, permitindo, assim, compará-los quantitativamente.

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 conforme os valores obtidos para os vetores após o treinamento do modelo. Assim, pode-se encontrar que a relação entre *big* e *bigger* é a mesma que entre *small* e *smaller*. Essa relação pode então ser reescrita através de uma operação vetorial como vetor("*big*") + vetor("*bigger*") - vetor("*small*") = vetor("*smaller*") (MIKOLOV *et al*., 2013).

Como exemplo de relação semântica, Mikolov et al. criaram tarefas para determinar as relações entre nomes de países e suas capitais. Assim podemos traduzir a relação "Paris está para França assim como Berlim está para Alemanha" como uma operação vetorial: 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 ao modelo CBOW e a outros modelos

de vetorização de palavras. 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 gasto para seu 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 arquiteturas comparadas precisaram de 14 dias de treino para criar o seu modelo. A redução no tempo de treinamento reduziu o custo computacional de forma que os modelos de vetorização de palavras se tornaram mais populares ainda (MIKOLOV *et al*., 2013).

Como anteriormente dito, *word embeddings* são capazes de capturar características semânticas e sintáticas, com base nisso, Hartmann (2016), os utiliza como *feature* para tarefas de similaridade semântica. Em seu trabalho é combinado um método já difundido, o TF-IDF (*term frequency–inverse document frequency*), com *word embeddings,* obtidos a partir de um corpus jornalístico do português brasileiro. O modelo apresentado pelo autor apresenta resultados superiores em comparação com o modelo base e o modelo TF-IDF em tarefas de similaridade semântica. (HARTMANN, 2016)

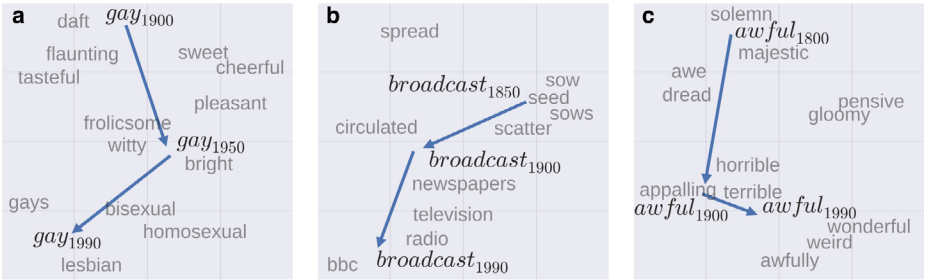
Tendo em mãos uma técnica não só capaz de capturar relações sintáticas e semânticas, como também de menor custo computacional, falta então a capacidade de se trabalhar com um *corpus* diacrônico. Ao propor leis estatísticas para a mudança semântica, Hamilton, Leskovec e Jurafsky (2016) utilizaram-se de vetores de palavras para obter seus resultados.

As leis propostas pelos autores são 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, pois 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).

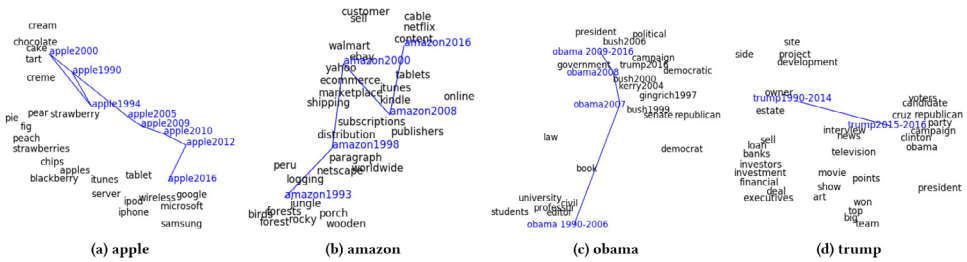
O estudo conseguiu, em um primeiro momento, validar mudanças de significado já conhecidas e foi em seguida usado para buscar as palavras que sofreram maior mudança semântica, aqui considerado o maior deslocamento no espaço vetorial ao longo dos períodos analisados (JURAFSKY; MARTIN, 2008).

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* diacrônicos é alinhar os eixos dos modelos. Devido às operações realizadas nos vetores para a obtenção de um modelo como o Skip-gram, as mesmas palavras 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




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.

## 4 Metodologia

A partir das visões do léxico em rede (como as de Givón e aquela proposta por Matoré), propõe-se, no presente trabalho, buscar a representação da mudança semântica de palavras relevantes para o português. Para isso, será utilizado o Corpus Histórico do Português Tycho Brahe, elaborado por De Sousa (2014).

Buscando uma metodologia similar à de Hamilton, Leskovec e Jurafsky (2016), o *corpus* foi dividido 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. Há ainda uma escassez de textos nos períodos mais antigos da língua, em especial no século XIV. 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 (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, buscaram-se as redes de relações para os seguintes termos: *homem, mulher*, *pai, mãe, terra* e *deus*. As palavras aqui mencionadas foram escolhidas por dois motivos: primeiramente, são centrais para a articulação de valores socioculturais, são termos geralmente carregados e podem evidenciar vieses; segundo, possuem frequência relativamente alta dentro do *corpus* e possibilitam 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 PLN existentes como Spacy e NLTK (Natural Language Toolkit). O *corpus* escolhido para a análise foi o *corpus* Anotado do Português Histórico Tycho Brahe (CTB), pioneiro no que concerne à língua portuguesa, que 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 séculos XIV e XX. Segundo De Sousa, as anotações realizadas nos textos têm como objetivo principal possibilitar a recuperação de informações filológicas e linguísticas dos textos (DE SOUSA, 2014).

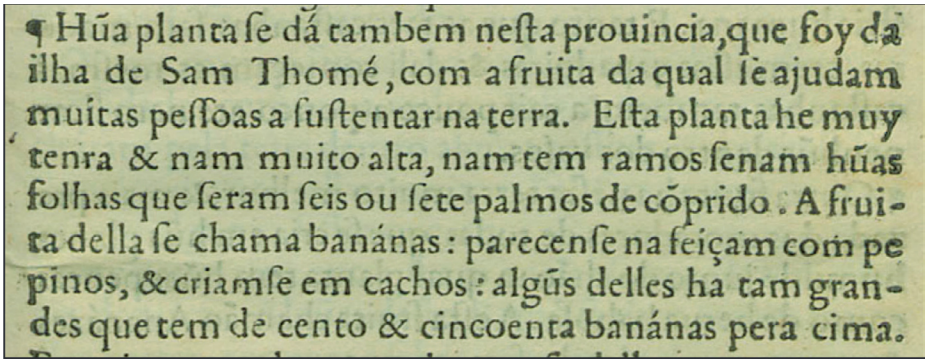
O *corpus* é composto de 88 textos, totalizando 3.544.628 palavras, sendo 58 textos anotados morfologicamente e 27 textos anotados sintaticamente. O *corpus* possui uma variedade de gêneros textuais, sendo eles: cartas, atas, textos narrativos, textos dissertativos, gramáticas, gazetas e jornais e textos 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 foi feito 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 informações relativas às decisões editoriais e à estrutura do texto (quebras de linha, parágrafos, seções, etc.) e também lida com intervenções interpretativas, como atualização grafemática, expansão de abreviaturas e atualização ortográfica. Na Figura 9, tem-se 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 fruita della ſe chama banánas: parecenſe na feiçam com pepinos, & criamſe em cachos: algũs delles ha tam grandes que tem de cento & cincoenta banánas pera cima.

Fonte: De Sousa (2014, p. 57).

A segunda etapa de anotação é a anotação morfossintática, que consiste na identificação e codificação das classes de palavras. A terceira e última etapa é a anotação sintática. Nessa etapa, é realizada a identificação e codificação da estrutura sintagmática da sentença, que foi realizada através de um parser sintático automático. Para realizar a anotação sintática do *corpus*, foi desenvolvido um parser sintático a partir do sistema Penn-Treebank. O parser 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, fez-se então uma análise exploratória preliminar do *corpus*. Para este trabalho, foram feitos recortes temporais nos textos do *corpus*. O agrupamento foi realizado a partir de uma adaptação da classificação proposta por Bechara (1985). A delimitação proposta por Bechara 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, em que os documentos escritos existentes são de variedade culta e erudita. Alguns dos fenômenos encontrados nessa variedade são:

- 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 destacados pelo autor são:

- 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.*

## 5 Análise exploratória

A partir desses conhecimentos, foi realizada uma análise exploratória do CBT com a intenção de avaliar quais termos atenderiam às 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[2];
- Padronização em caixa baixa;
- Remoção de s*topwords*;
- Remoção de acentuação.

O texto foi inicialmente tokenizado e, em seguida, foi padronizado completamente em caixa baixa, evitando assim que tokens iguais sejam considerados diferentes devido à escrita em maiúscula. Dessa forma, os tokens "Portanto" e "portanto" correspondem ao mesmo token, "portanto".

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

A retirada de *stopwords* é um processo comum em tarefas de PLN. As *stopwords* são palavras gramaticais ou consideradas pouco relevantes semanticamente, como pronomes, preposições e artigos. Podem também ser palavras filtradas por não serem de interesse do trabalho em questão. Sendo assim, frequentemente elas são ignoradas em análises de PLN. Existem diversas listas de Stopwords disponíveis com diferentes critérios, e neste trabalho foi usada a lista fornecida por padrão pela biblioteca NLTK (BIRD; KLEIN; LOPER, 2009).

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

Tabela 1 - Dez 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 |

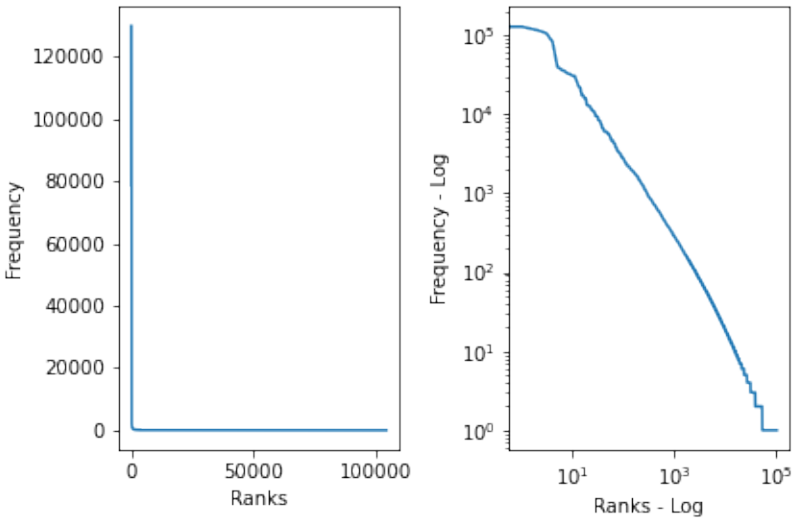
Fonte: Elaboração própria.

Na Figura 10 à esquerda, vê-se a distribuição de frequências para as palavras do *corpus*. Aqui, se vê como poucos termos possuem uma frequência alta, enquanto a grande maioria do *corpus* possui uma frequência similar, mais baixa. Já o gráfico à direita mostra o mesmo gráfico em escala logarítmica, em que se tem uma relação linear entre as potências das frequências e seu posicionamento na ordem de frequência.

A Figura 10 é um exemplo da Lei de Zipf. A Lei de Zipf é uma constatação empírica que nos diz que, a posição, ou ranque, de uma palavra em uma tabela de frequências ordenada de forma decrescente, é inversamente proporcional a sua frequência no *corpus* analisado. Assim, a segunda palavra mais frequente em um *corpus* possui frequência aproximadamente duas vezes menor que a palavra mais frequente e assim

por diante. Em geral, as palavras que aparecem com maior frequência são palavras de função gramatical.

Figura 10 - Distribuição de frequências das palavras do corpus.




Fonte: Elaboração própria.

Por fim, foram encontradas as frequências e o ranque, após a retirada de *stopwords*, para as palavras selecionadas para a análise, indicadas na Tabela 2.

Tabela 2 - Frequência das palavras a serem analisadas e seu ranque

| Palavra | Frequência | Ranque |
|---|---|---|
| deus | 4602 | 3 |
| homem | 2506 | 32 |
| terra | 2209 | 45 |
| pai | 1312 | 133 |
| mulher | 1229 | 145 |
| mãe | 798 | 240 |

Fonte: Elaboração própria.

Para as formas “pai”, “mae”, “deus”, “homem”, “mulher” e “terra” foram encontradas palavras de frequência consideradas satisfatórias e

que possivelmente carregam vieses em seus contextos de uso, portanto foram elas as escolhidas para serem analisadas.

A partir de uma inspeção visual inicial, constatou-se que os números de textos presentes nas fases Arcaica e Arcaica Média aparentavam ser muito menores que os números de textos nas demais fases. Assim, já em um primeiro recorte temporal do *corpus* foram unidas as fases Arcaica e Arcaica Média em um só período de tempo, aqui denominado Período I. Vê-se na Tabela 3 que o número de *tokens* para o Período I, mesmo constituindo-se da união de dois outros períodos, ainda é baixo em comparação com os demais períodos. Foi mantida essa separação apesar do desbalanceamento temporal aqui encontrado.

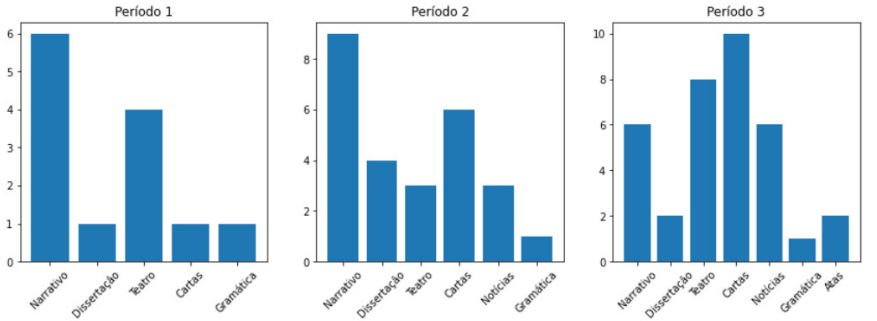
Tabela 3 - Fases do português (adaptado de Bechara (1985)) e respectivo número de *tokens* no *corpus* Tycho Brahe.

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

Fonte: Elaboração própria.

Com o intuito de se analisar o balanceamento do *corpus* para os diferentes gêneros textuais encontrados, obtém-se a Figura 11, que indica o número de textos para cada gênero textual.

Figura 11 - Gêneros textuais presentes




Fonte: Elaboração própria.

Após serem analisados os balanceamentos do *corpus* e agrupados os textos em recortes temporais, foi feita a limpeza e processamento do *corpus* para a posterior criação dos modelos *Skip-gram*.

## 6 Limpeza e treinamento dos modelos

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 PLN e consiste em deflexionar uma palavra, retornando-a para sua forma base, dicionarizada. O resultado desse processo são nomes no singular e no masculino, e verbos no infinitivo. 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".

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

Todos os passos de limpeza e pré-processamento foram feitos através da biblioteca Spacy[3], baseada na linguagem de programação Python em sua versão 3.1.1.

Os modelos Skip-gram foram treinados, um para cada período, através do código fornecido por Mikolov.[4]

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 4.

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: Elaboração própria.

O parâmetro *size* diz respeito ao tamanho do vetor, ou número de dimensões que o vetor de cada palavras possuirá após o treinamento. O

[3] Disponível em: https://spacy.io/

[4] Disponível em: https://github.com/tmikolov/word2vec

parâmetro *window* se refere à janela de treinamento: o seu valor determina o número de *tokens* antes e 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 ao se substituir uma palavra em exemplo por uma palavra aleatória do *corpus*. Pode-se ver um exemplo negativo na Figura 12, onde a forma *apricot* é emparelhada em contextos reais na coluna da esquerda e em contextos não existentes no *corpus* na coluna da direita. O valor fornecido corresponde à razão entre o número de amostras negativas e positivas: neste caso, tem-se 25 vezes mais amostras negativas que positivas. Por fim, o parâmetro *iter* diz respeito ao número de iterações necessárias para o treinamento: neste caso, o treinamento foi repetido 25 vezes.

Figura 12 - Exemplos positivos e negativos no corpus

| Exemplos positivos + | | **Exemplos Negativos -** | |
|---|---|---|---|
| Palavra (n) | Contexto (n+1) | Palavra (n) | Contexto (n+1) |
| deus | ser | deus | cachorro |
| deus | governar | deus | sete |
| deus | irado | deus | meu |

Fonte: Elaboração própria

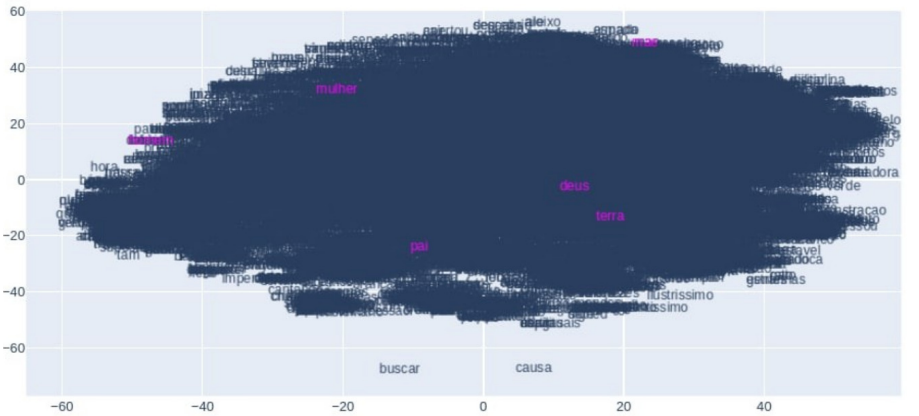
Após a obtenção dos vetores de palavras, resta apenas gerar a visualização gráfica.

## 7 Visualizações de redes de relações semânticas

Assim como Osgood et al. representam palavras em três dimensões, os modelos de vetorização de palavras aqui utilizados geram vetores com uma dimensão definida no momento do treinamento. Para este trabalho foram criados vetores de 300 dimensões e, por isso, deve-se reduzir essas dimensões a apenas duas para que possam ser visualizadas em um espaço bidimensional. Para isso, foi utilizada a biblioteca Gensim 4.0.1 para Python.

Com ela, foi usado o 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 13. As imagens finais foram obtidas após a devida ampliação da imagem.

Figura 13 - Exemplo de representação gráfica da rede de relações semânticas

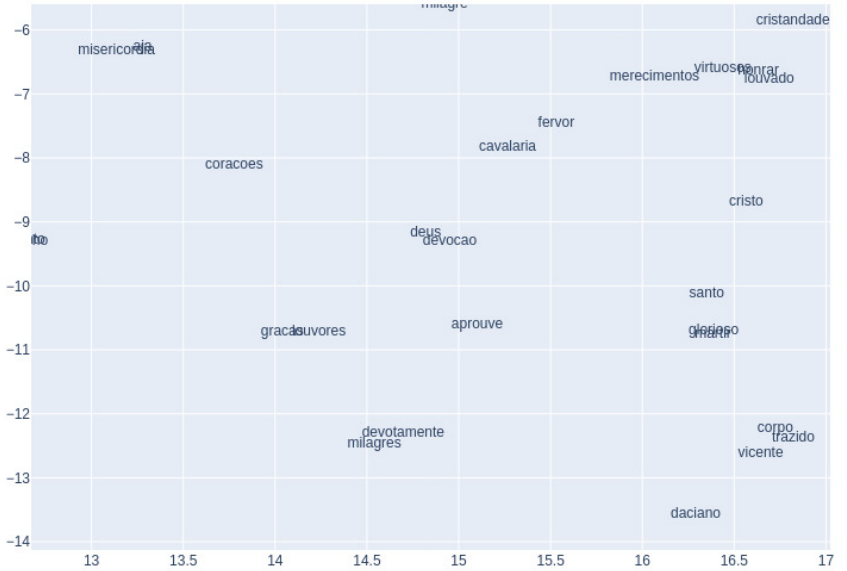



Fonte: Elaboração própria.

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

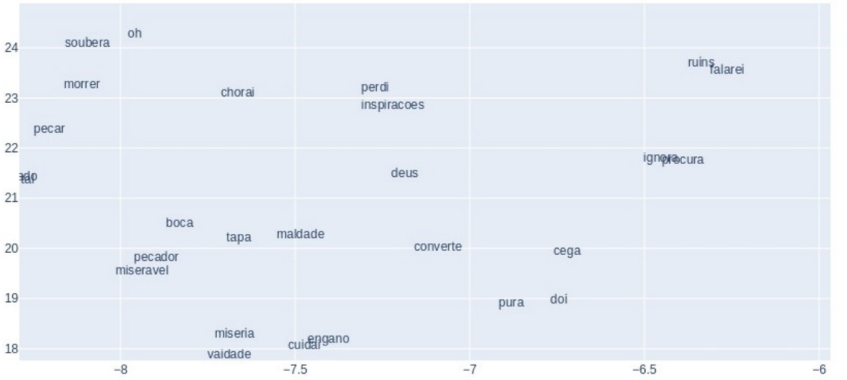
Primeiramente, o gráfico gerado para a palavra "deus" no período I mostra um contexto similar ao esperado, com palavra como "devocao", "louvores", "devotamente" (Figura 14), em sua proximidade. Já para o período II, vê-se palavras como "inspiracoes", "converte", "perdi" e "maldade" (Figura 15) próximas. Por fim, no período III têm-se as formas "amas", "aflitos", "misericordia", "divinos", "preconceito", "santissima" (Figura 16).

Figura 14 - Rede de relações semânticas da palavra “deus”, período I



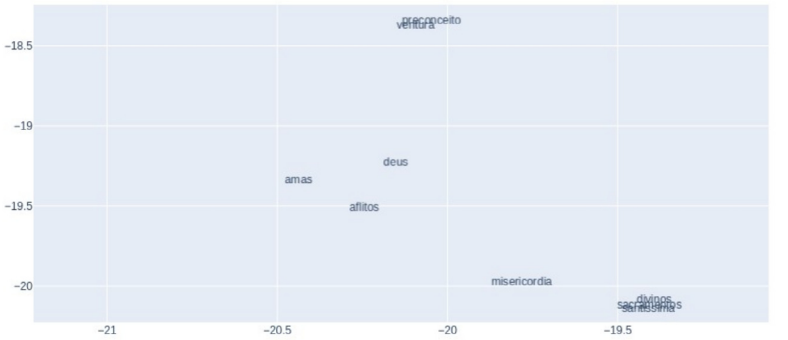

Fonte: Elaboração própria.

Figura 15 - Rede de relações semânticas da palavra deus, período II




Fonte: Elaboração própria.

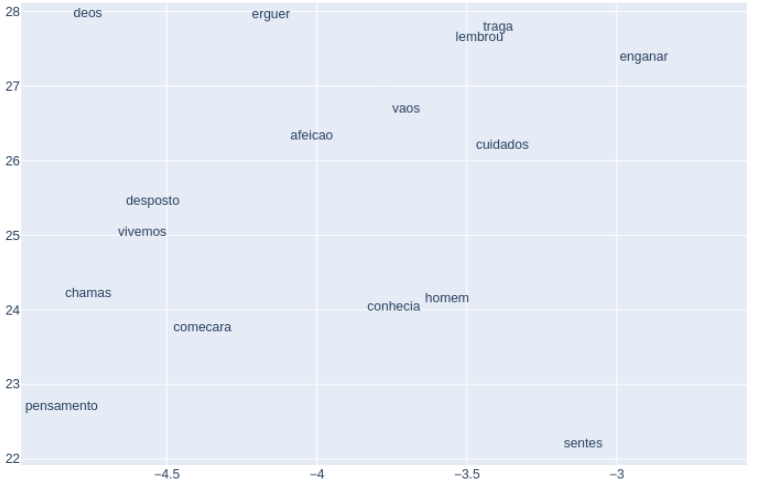
Figura 16 - Rede de relações semânticas da palavra “deus”, período III




Fonte: Elaboração própria

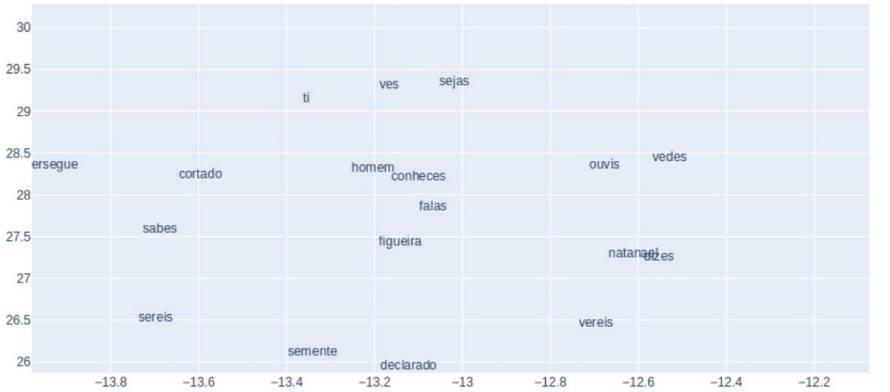
Agora, as relações semânticas para a forma “homem”, no período I vemos as palavras “conhecia”, “cuidados”, “sentes”, “comecara” mais próximas (Figura 17). Já para o período II têm-se as palavras “conheces”, “falas”, “figueira”, “ouvis”, “ves”, “sejas” (Figura 18). No Período III vê-se as palavras “miseravel”, “ateu”, “ímpio”, “solene”, “livrar” (Figura 19).

Figura 17 - Rede de relações semânticas da palavra “homem”, período I



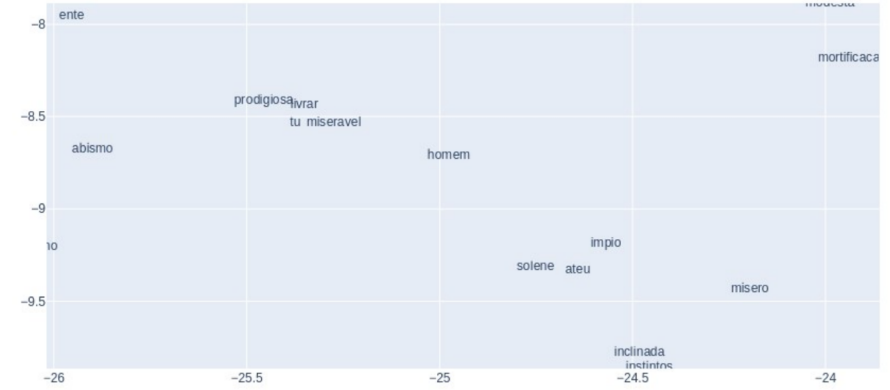

Fonte: Elaboração própria.

Figura 18 - Rede de relações semânticas da palavra “homem”, período II




Fonte: Elaboração própria.

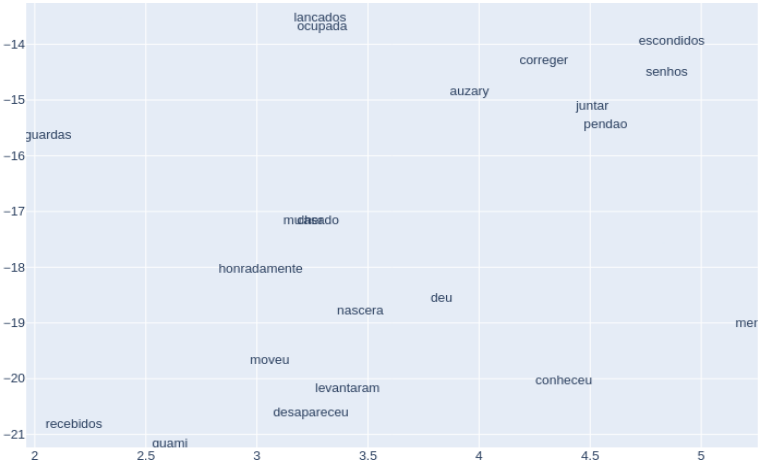
Figura 19 - Rede de relações semânticas da palavra “homem”, período III




Fonte: Elaboração própria.

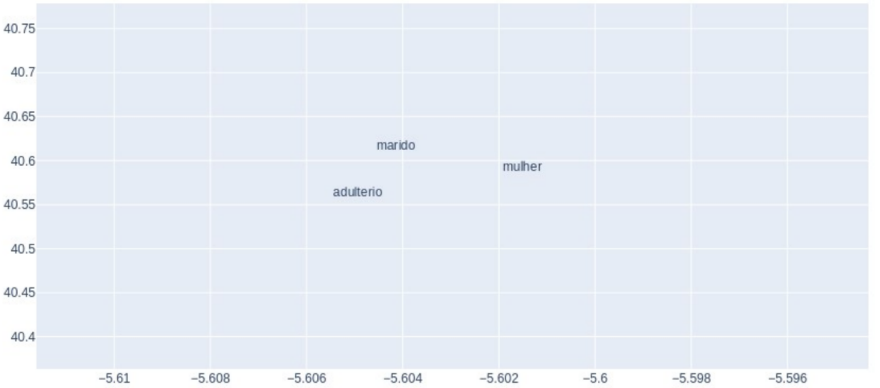
Para a rede da palavra “mulher”, vemos no período I, que ela ficou bastante próxima da palavra “casado”, seguida das palavras “nascera”, “deu”, “nascera”, “honradamente” (Figura 20). Já no período II, temos as palavras “marido” e “adultério” muito próximas (Figura 21). Finalmente no período III, temos as palavras “desgraçada”, “coitadinha”, “marido” e “margarida” na proximidade da palavra analisada (Figura 22).

Figura 20 - Rede de relações semânticas da palavra “mulher”, período I



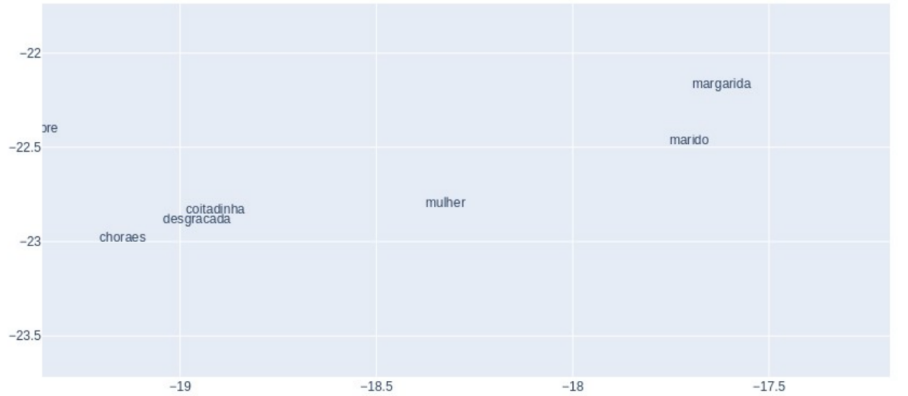

Fonte: Elaboração própria

Figura 21 - Rede de relações semânticas da palavra “mulher”, período II




Fonte: Elaboração própria.

Figura 22 - Rede de relações semânticas da palavra “mulher”, período III



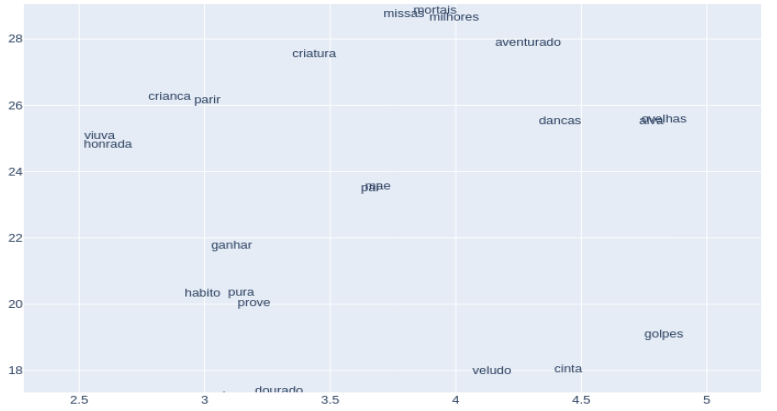

Fonte: Elaboração própria.

As formas “pai” e “mae” foram analisadas em conjunto devido à proximidade que elas se encontram nos resultados do período I. Nele, elas se sobrepõem, como mostrado na Figura 23. Temos as palavras “criatura”, “ganhar”, “cinta”, “pura”, “missas”, “parir”, “crianca”, “viuva” 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 24. Já a palavra “mae” se encontra próxima de expressões como “casados”, “bodas”, “legitimo”, “casal”, “embaracos” e “ajustada”, como visto na Figura 25.

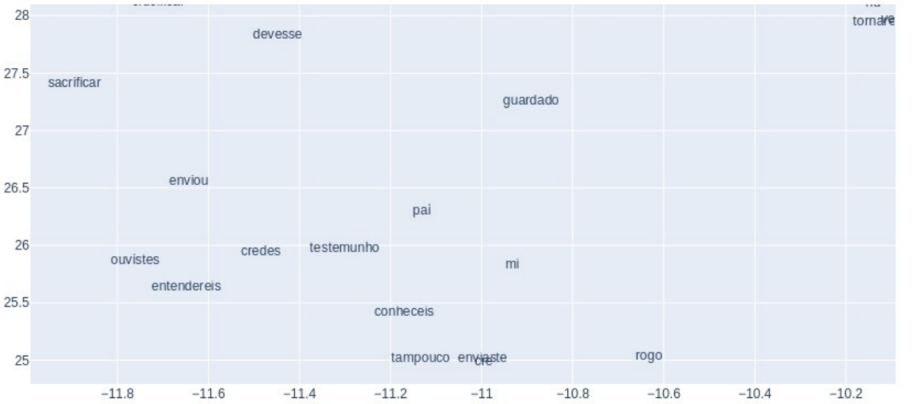
No período III, a rede da palavra “pai” mostra proximidade com “consolar”, “paterno”, “filho”, além do verbo “tourear”, Figura 26. Já o campo de “mae”, mostra as palavras “virtuosa”, “filha”, “carinhosa”, vide Figura 27.

Figura 23 - Rede de relações semânticas das palavras “pai” e “mae”, período I.



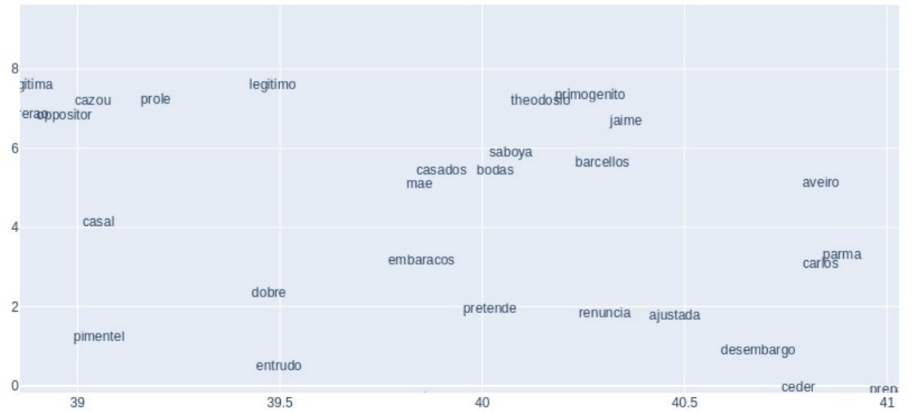

Fonte: Elaboração própria.

Figura 24 - Rede de relações semânticas da palavra “pai”, período II



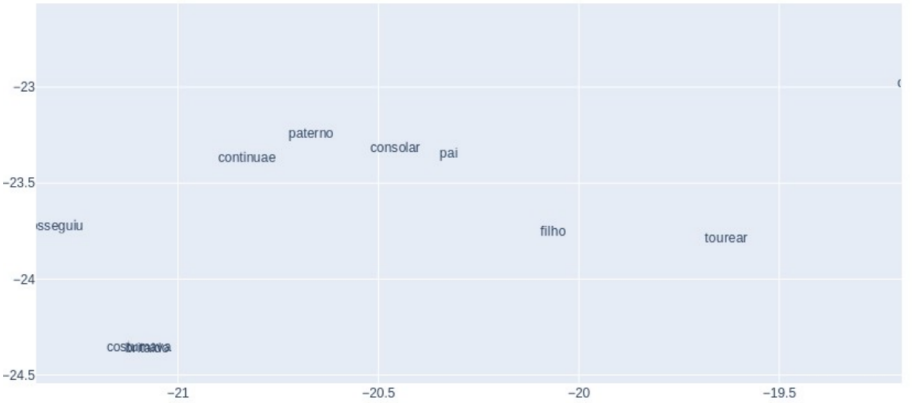

Fonte: Elaboração própria.

Figura 25 - Rede de relações semânticas da palavra “mae”, período II



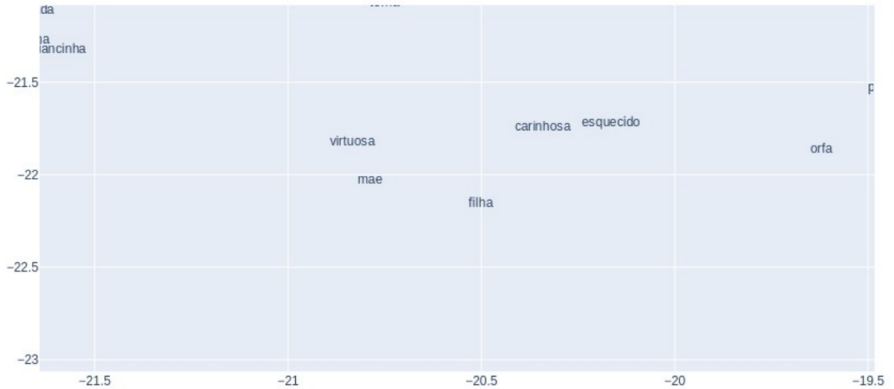

Fonte: Elaboração própria

Figura 26 - Rede de relações semânticas da palavra “pai”, período III




Fonte: Elaboração própria.

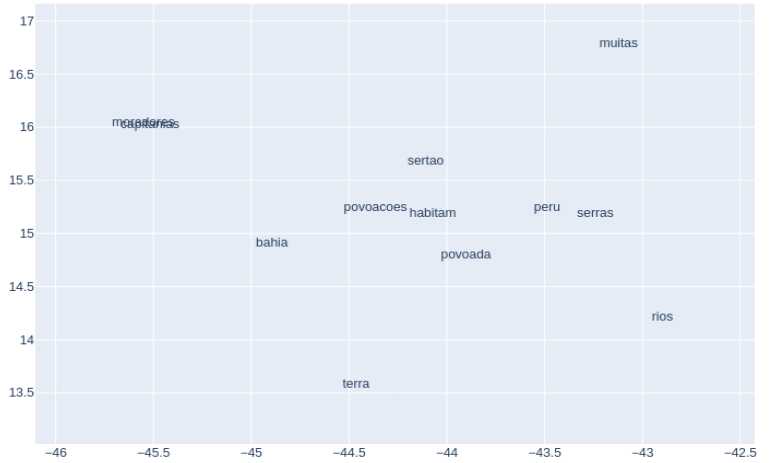
Figura 27 - Rede de relações semânticas da palavra “mae”, período III




Fonte: Elaboração própria.

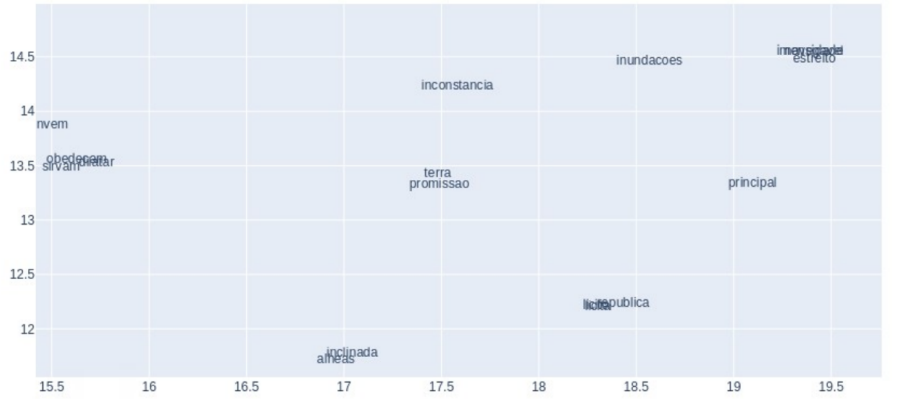
Para o período I, a palavra “terra” possui em sua vizinhança as palavras “sertao”, “habitam”, “povoada”, “bahia”, “rios”, conforme indicado na Figura 28. Já para o período II a Figura 29 mostra as palavras “inconstancia”, “promissao” e destaca-se a palavra “republica”. Finalmente o período III apresenta palavras como “campinas”, “montanhas”, “ribeiras”, “ventos” e “tempestades”, como visto na Figura 30.

Figura 28 - Rede de relações semânticas de “terra”, período I



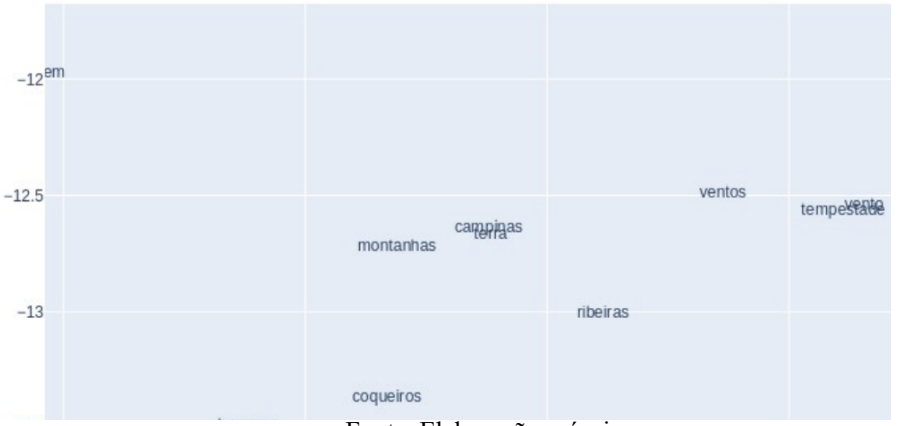

Fonte: Elaboração própria

Figura 29 - Rede de relações semânticas de “terra”, período II




Fonte: Elaboração própria

Figura 30 - Rede de relações semânticas para “terra”, período III




Fonte: Elaboração própria.

## 8 Discussão dos resultados

Inicialmente, vê-se que as redes de relações passaram por mudanças ao longo dos períodos analisados. Observa-se que a palavra “deus” 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 observa-se essa palavra associada a verbos sensoriais como “vedes”, “ouvis”, “falas”. Essa ausência de palavras fortemente ligadas à forma “homem” pode ser uma consequência de seu uso como forma de indeterminação do sujeito nesses períodos, como informado por Bechara (1985). Já no último momento, vê-se a forma “homem” próxima a palavras de viés negativo, como “miseravel” e “impio”, e também à palavra “ateu” e à 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 “adulterio”. 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".

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 "mae", no mesmo período, mostra palavras relacionadas a família e casamento, como "casados", "casal", "prole", "bodas", "primogenito". 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 "orfa" em sua proximidade. Também se vê 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 "republica". 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 no da 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 dos períodos analisados.

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,

porém, 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 nominais apresentando o gênero feminino e flexão de grau como em "coitadinha", "desgraçada", enquanto eram esperadas suas formas dicionarizadas. O 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", já que o modelo utiliza um conjunto de regras para gerar os lemas.

## 9 Conclusão

O presente trabalho buscou analisar palavras significativas e os diferentes contextos semânticos em que elas surgem em diferentes períodos de um *corpus* diacrônico. A análise foi realizada por meio do uso da técnica de PLN conhecida como *word 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 precisa. Esse processo é de extrema importância para línguas de morfologia rica, como o português, e não recebe tanta atenção devido aos sistemas de PLN.

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 por meio do uso de um *corpus* que possui mais de 410 milhões de tokens, totalizando mais de 100 milhões de tokens por período estudado – um valor muito distante da quantidade de tokens obtidos com o uso do *corpus* Tycho Brahe. Apesar de os autores citarem outras formas de obtenção de *word embeddings* e, também, recomendarem 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, quanto textos jornalísticos. Além disso, esses diferentes gêneros encontram-se desbalanceados quanto à sua representação no *corpus*.

Considerando as limitações encontradas nesta análise, propõe-se para estudos futuros o uso de formas alternativas de obtenção de *word embeddings*, comparando os resultados com os aqui obtidos. Além disso, sugere-se o uso de outro lematizador, que possa fornecer um resultado mais 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 os autores constataram, uma análise quantitativa que utilize metodologia similar para dados diacrônicos nessa língua.

## Agradecimentos

Os autores agradecem à Profa. Adriana Pagano (FALE/UFMG) e ao Prof. Bruno Rocha (FALE/UFMG) pela participação na banca de avaliação da monografia que levou a este artigo e pelas valiosas sugestões ao trabalho.

## Declaração de autoria

Ambos os autores conceberam e planejaram o estudo. Lucas F. Lage realizou o pré-processamento dos dados, gerou as análises, interpretou os resultados e escreveu o texto. Evandro L. T. P. Cunha orientou a realização do trabalho, auxiliou na interpretação dos resultados e revisou o texto. Ambos aprovaram a versão final.

## Referências

BĂLAN, O. *et al.* Emotion classification based on biophysical signals and machine learning techniques. *Symmetry*, [s.l.], v. 12, n. 1, 2020. DOI: <https://doi.org/10.3390/sym12010021>. Acesso em: 31 jan. 2022.

BECHARA, E. *As fases históricas da língua portuguesa:* tentativa de proposta de nova periodização. Niterói: 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*, [s.l.], v. 11, n. 101, 2014. DOI: <http://dx.doi.org/10.1098/rsif.2014.0841>. Acesso em: 31 jan. 2022

CAMBRAIA, C. N. Da lexicologia social a uma lexicologia sócio-histórica: caminhos possíveis. *Revista de Estudos da Linguagem*, Belo Horizonte, v. 21, n. 1, p. 157–188, 2013. DOI: <http://dx.doi.org/10.17851/2237-2083.21.1.157-188>. Acesso em: 31 jan. 2022.

CUNHA, A. F. DA. Funcionalismo. *In*: MARTELOTTA, M.E. (org.). *Manual de linguística*. São Paulo: Contexto, 2008. p. 157–176.

DE SOUSA, M. C. P. O *corpus* Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa*, São Paulo, v. 16, n. esp., p. 53–93, 2014. DOI: https://doi.org/10.11606/issn.2176-9419.v16ispep53-93

DORES, M. V. P. das; TOLEDO, C. V. S. De "lepra" a "hanseníase": uma análise lexicológica de base sócio-histórica. *Diacrítica*, [s.l.], v. 32, n. 1, p. 179–208, 2018.

FIRTH, J. R. A synopsis of linguistic theory, 1930-1955. *In*: FIRTH, J. R. *Studies in linguistic analysis*. Oxford: Blackwell, 1957. p. 1-32.

GALVES, C.; ANDRADE, A. L.; FARIA, P. *Tycho Brahe Parsed Corpus of Historical Portuguese*. Disponível em <http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip>. Acesso em: dez. 2017.

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

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

GRIFFIN, C. *Graph Theory*: Penn State Math 485 Lecture Notes. 2017. Disponível em: <https://www.personal.psu.edu/cxg286/Math485.pdf>. Acesso em: 22 fev. 2022.

HAMILTON, W. L.; LESKOVEC, J.; JURAFSKY, D. Diachronic word embeddings reveal statistical laws of semantic change. *In*: 1ST INTERNATIONAL CONFERENCE ON LEARNING

REPRESENTATIONS, 2016. Disponível em: <https://arxiv.org/abs/1605.09096>. Acesso em: 31 jan. 2022.

HARTMANN, N. S. Solo queue at ASSIN: Combinando abordagens tradicionais e emergentes. *Linguamática*, [s.l.], v. 8, n. 2, p. 59-64, 2016

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

LABOV, W. Some principles of linguistic methodology. *Language in Society*, [s.l.], v. 1, n. 1, p. 97–120, 1972. DOI: <https://doi.org/10.1017/S0047404500006576>

MATORÉ, G. La lexicologie sociale. *L'Information Littéraire*, Paris, n. 2, mar./abr. 1949.

MICHEL, J. B. *et al.* Quantitative analysis of culture using millions of digitized books. *Science,* [s.l.], v. 331, n. 6014, p. 176–182, 2011. DOI: <10.1126/science.1199644> Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279742/>. Acesso em: 31 jan. 2022.

MIKOLOV, T. *et al*. Efficient estimation of word representations in vector space. *In*: 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.

RAFAEL, G. C. R. A.; SIMIÃO, D. P. Aidético e soropositivo: análise sócio-histórica da concorrência entre qualificadores utilizados em referência a portadores do HIV. *Inventário*, n. 23, p. 45–68, 2019.

ROBIN, R.; DE MENESES BOLLE, A. B. *História e lingüística.* São Paulo: Editora Cultrix, 1977.

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. DOI: https://doi.org/10.1016/S0022-5371(79)90355-4

ZIPF, G. K. *Human behavior and the principle of least effort:* An introduction to human ecology. [s.l.] Books, 2016.