

Identificação e caracterização da Mudança Semântica Lexical: uma análise na língua portuguesa

Identification and characterization of Lexical Semantic Change:
an analysis in the Portuguese language

Laila Mota  
FORMAS Research Center
Instituto de Computação
Universidade Federal da Bahia

Eloize R. Marques Seno  
Área de Computação
Instituto Federal de São Paulo

Rerisson Cavalcante  
FORMAS Research Center
Instituto de Letras
Universidade Federal da Bahia

Daniela Barreiro Claro  
FORMAS Research Center
Instituto de Computação
Universidade Federal da Bahia

Resumo

Este estudo apresenta uma metodologia comparativa para a identificação e a caracterização da Mudança Semântica Lexical (MSL) em língua portuguesa, integrando Modelos de Linguagem Estáticos e Modelos de Linguagem Contextuais, ajustados (*fine-tuned*) em corpora diacrônicos (ALiB-Twitter e Tycholina). Os resultados demonstram que, embora os modelos estáticos (Word2Vec/FastText) sejam eficazes na quantificação de transformações globais, sua sensibilidade é limitada por *corpora* ruidosos e pela dificuldade de representar a polissemia. Em contraste, a análise com modelos contextuais, suportada por algoritmo de agrupamento, revelou-se capaz de capturar mudanças sutis de uso, caracterizando a pejoração e a ampliação de sentido, com uma validação humana qualitativa dos agrupamentos gerados.

Palavras chave

modelos de linguagem; mudança semântica lexical; *embeddings*; espaço semântico

Abstract

This study describes a comparative methodology for identifying and characterizing Lexical Semantic Change (LSC) in Portuguese, integrating both Static Language Models and Contextual Language Models, fine-tuned on diachronic *corpora* (ALiB-Twitter and Tycholina). Our findings describe that although static models (Word2Vec/FastText) are effective for quantifying global transformations, their sensitivity is limited by noisy *corpora* and by the challenge of representing polysemy. On the other hand, contextual models are capable of capturing subtle usage shifts, characterizing pejoration and sense broadening, with qualitative human validation.

Keywords

language models; lexical semantic change; *embeddings*; semantic space

1. Introdução

A língua é dinâmica e está em constante transformação, refletindo as mudanças culturais, sociais e tecnológicas da sociedade que a utiliza (Aitchison, 2002). O estudo dessas mudanças linguísticas visa identificar, interpretar e avaliar as alterações no léxico, também conhecidas como Mudança Semântica Lexical (MSL). Além de revelar transformações, a análise da MSL oferece informações sobre a história da língua e sua conexão com a cultura e a sociedade. Essa análise é tradicionalmente estudada por meio de análises manuais de *corpora* e dicionários, limitadas em volume, gêneros e períodos de tempo analisados. Entretanto, o uso de ferramentas computacionais nos últimos anos têm sido cada vez mais frequentes para automatização destes (Beck, 2020; Kutuzov & Giulianelli, 2020; Martinc et al., 2020).

Mudanças semânticas no léxico envolvem o desaparecimento, surgimento ou alteração de significados das palavras. Segundo Crowley & Bowern (2010), essas mudanças podem ser classificadas em seis tipos: melhora, pejoração, ampliação, estreitamento, bifurcação e mudança. Essas modificações podem ser observadas em todos os idiomas, a exemplo do caso do termo “cloud” (nuvem) que, há duas décadas, era associado principalmente ao clima. Hoje, seu significado está mais ligado ao armazenamento remoto de dados (Barranco et al., 2018). Neste

Tipo de mudança	Descrição
Melhora	palavras cujo significado muda para ser mais positivo
Pejoração	palavras cujo significado muda para ser mais negativo
Ampliação	adição de sentidos àqueles que expressava originalmente, preservando os sentidos originais
Estreitamento	uma palavra passa a se referir apenas a parte do significado original
Bifurcação	uma palavra adquire outro significado que se relaciona de alguma forma com o significado original
Mudança	uma palavra perde completamente seu significado original e adquire um novo significado

Tabela 1: Tipos de mudança semântica segundo Crowley & Bower (2010).

trabalho buscamos identificar segundo as categorias elencadas por Crowley & Bower (2010) (Tabela 1) a MSL em unidades lexicais da língua portuguesa.

Em relação aos *corpora*, o estudo da MSL pode utilizar *corpora* históricos, analisando quantitativamente a frequência de palavras ao longo do tempo ou qualitativamente o contexto de seu uso (Hilpert & Gries, 2016; Silva, 2018). Dicionários de diferentes épocas também são recursos valiosos (Ruas, 2023) utilizados neste tipo de investigação. Com o avanço dos modelos de linguagem, análises de MSL foram aplicadas principalmente em *corpora* de língua inglesa, evidenciando uma lacuna na aplicação desses modelos em outras línguas, como o português. A utilização de modelos de linguagem requer um conjunto de *corpora* que permita o treinamento de modelos com o intuito de obter uma generalização para a tarefa que se propõe. Não se tem conhecimento de um *corpus* anotado com os anos de publicação dos textos que aborde longos períodos (centenas de anos) com grandes volumes de dados para a língua portuguesa. Assim, com o intuito de dirimir essa lacuna, este trabalho utilizou dois *corpora*, o *corpus* histórico do português Tycho Brahe (Galves, 2018) e o *corpus* Carolina (Crespo et al., 2023), e anotou manualmente as divisões temporais, se caracterizando assim como uma primeira contribuição desta pesquisa, denominado a partir de então de *corpus Tycolina*.

Considerando a evolução do significado ao longo do tempo, a análise da semântica diacrônica enfrenta desafios intrínsecos à linguagem. Dentre esses desafios, esta pesquisa destaca a polissemia, em que uma palavra possui múltiplos sentidos em um mesmo espaço de tempo, dificultando a identificação precisa de suas mudanças semânticas (De Saussure, 2006). Esse processo é gradual e sutil, o que torna difícil determinar exatamente quando e como uma mu-

dança ocorre (Coelho et al., 2015). Além disso, o grande volume de dados textuais, abrangendo diversos períodos e gêneros, torna a análise manual complexa e demorada. Nesse contexto, o uso de métodos computacionais eficientes pode ser uma abordagem para esta análise.

O presente trabalho descreve uma metodologia para analisar as mudanças semânticas lexicais comparando desempenho de modelos de linguagem estáticos (Word2vec (Mikolov et al., 2013a) e Fasttext (Bojanowski et al. (2017a); Joulin et al. (2017a))) e contextuais (BERTimbau (Souza et al., 2020) e XLM-Roberta (Conneau et al., 2020)), sob a perspectiva diacrônica do desaparecimento, surgimento ou alteração do significado das palavras, sendo as alterações classificadas em seis tipos, conforme proposto por Crowley & Bower (2010) (vide Tabela 1). A investigação busca responder duas questões de pesquisa:

- Qual é a eficiência dos modelos estáticos em identificar mudanças semânticas (*semantic shift*)?
- De que forma os modelos contextuais permitem caracterizar a mudança semântica (*semantic drift*)?

Os experimentos foram divididos em dois grupos: experimentos com modelos de linguagem estáticos (Word2vec, Fasttext) e experimentos com modelos de linguagem contextuais (BERTimbau, XLM Roberta). Os experimentos empregaram dois *corpora* (ALiB-Twitter e Tycolina) para geração de representações vetoriais das palavras (*embeddings*) e avaliação da MSL das unidades lexicais selecionadas.

Assim, os modelos desenvolvidos através deste trabalho com o intuito de analisar as mudanças semânticas lexicais se caracterizam como a segunda contribuição desta pesquisa. Os resultados obtidos demonstraram a capacidade dos modelos

em capturar e representar a MSL, permitindo caracterizar as mudanças semânticas em diferentes janelas temporais para a língua portuguesa.

Este trabalho está organizado em seções como segue: a seção 2 posiciona o trabalho no estado da arte. A seção 3 descreve o método desenvolvido neste trabalho, a seção 4 apresenta os experimentos, a seção 5 descreve os resultados dos experimentos e as discussões são apresentadas na seção 6. Por fim, a seção 7 conclui e apresenta assuntos perspectivas de trabalhos futuros.

2. Trabalhos Relacionados

Duas grandes abordagens, quanto ao tipo do modelo, são popularmente utilizadas em trabalhos relacionados: a abordagem baseada em análises estatísticas e *embeddings* estáticos (Baranco et al., 2018; Hamilton et al., 2016a; Dubossarsky et al., 2019; Fonseca Lage & Teixeira Paradelo Cunha, 2022) e a abordagem baseada em *embeddings* contextuais (Beck, 2020; Kutuzov & Giulianelli, 2020; Martinc et al., 2020).

Em relação aos *embeddings* estáticos, Baranco et al. (2018) apresentaram uma abordagem para inclusão de aspectos temporais nos espaços de representação por meio de um modelo de espaço vetorial reflexivo do tempo para dados de texto temporal capaz de capturar mudanças de curto e longo prazo no significado das palavras. Em Hamilton et al. (2016a) foi desenvolvida uma metodologia para quantificar a mudança semântica de palavras através da avaliação dos seus *embeddings*. A partir dos resultados, os autores propuseram duas leis de mudança semântica: a Lei da Conformidade (que afirma que a taxa de mudança semântica de uma palavra é inversamente proporcional à sua frequência) e a Lei da Inovação (independente da frequência da palavra, as palavras com mais significados, ou seja, polissêmicas, têm maiores taxas de mudança). Em Dubossarsky et al. (2019) os autores investigaram a Referência Temporal (*Temporal Referencing*, TR), um método para modelar a MSL que evita o alinhamento de espaços vetoriais ao incluir um token temporal específico no *corpus* de treinamento. Fonseca Lage & Teixeira Paradelo Cunha (2022) investigaram a viabilidade do uso de modelos de linguagem para análise de mudança semântica na língua portuguesa, utilizando a estratégia Skip-gram do algoritmo Word2vec.

Considerando os *embeddings* contextuais, Beck (2020) apresentou o sistema DiaSense, desenvolvido para a Task 1 do SemEval-2020 (Schlechtweg et al., 2020). O DiaSense utiliza *embeddings* de palavras contextualizadas, gerados

pelo modelo BERT pré treinado, para modelar as mudanças de sentido das palavras ao longo do tempo. Kutuzov & Giulianelli (2020) também participaram da Task 1 do SemEval-2020, focando na sub tarefa que envolve a classificação de palavras com base no grau de mudança semântica ao longo do tempo. Os autores avaliaram que *embeddings* contextualizados fornecem uma correlação forte com julgamentos humanos sobre mudanças semânticas, superando *embeddings* estáticos.

Diferentemente de outros estudos voltados ao português, como o de Fonseca Lage & Teixeira Paradelo Cunha (2022), que se restringe à avaliação das mudanças semânticas utilizando uma única arquitetura (Word2Vec), esta pesquisa apresenta uma contribuição metodológica ao propor uma abordagem comparativa empregando modelos de linguagem estáticos (Word2Vec e FastText) e contextuais (BERTimbau e XLM-RoBERTa), na análise da Mudança Semântica Lexical em unidades lexicais da língua portuguesa, utilizando a representação do espaço semântico e a vizinhança de unidades lexicais com modelos estáticos e a representação do espaço semântico e agrupamento de sentidos com modelos contextualizados. O estudo também avança na escala dos dados ao compilar dois *corpora* (*corpus* histórico do português Tycho Brahe e *corpus* Carolina), ampliando a cobertura diacrônica.

Embora esses modelos sejam populares em tarefas de Processamento de Linguagem Natural para o inglês, sua aplicação na análise diacrônica de itens lexicais do português ainda é limitada. Adicionalmente, enquanto a maioria dos estudos foca em um único modelo e em termos com mudanças semânticas já conhecidas, este trabalho busca analisar, por meio de técnicas de representação vetorial, a mudança de sentido de unidades lexicais selecionadas com base em critérios específicos. A análise é realizada ao longo do tempo, a partir das coocorrências de lexias em diferentes épocas e da amostragem de agrupamentos de sentidos.

2.1. Shift vs Drift semântico

Os trabalhos nesta área de pesquisa utilizam diversos termos, majoritariamente em inglês, para descrever o fenômeno da Mudança Semântica Lexical (MSL), sendo os mais populares *semantic shift* e *semantic drift*. Embora frequentemente empregados de forma intercambiável para representar a alteração diacrônica no significado de uma unidade lexical, a distinção entre os termos é

metodologicamente relevante para o presente estudo, alinhando-se as abordagens propostas em trabalhos recentes (de Sá et al., 2026).

Neste trabalho, adotamos uma distinção entre os termos: o termo *semantic shift* é definido como a detecção de uma mudança semântica ocorrida entre os períodos T_1 e T_2 , mensurada por métricas globais como a Similaridade de Cosseno. Por outro lado, o termo *semantic drift* refere-se à análise qualitativa e tipológica da natureza dessa mudança, abrangendo a caracterização dos fenômenos. O presente trabalho analisa ambos os fenômenos (*semantic shift* e *semantic drift*) nos corpora ALiB-Twitter e Tycolina.

3. Método de Análise da MSL em PT

O método empregado neste trabalho para a análise da mudança semântica lexical (MSL) em unidades lexicais da língua portuguesa foi estruturado em cinco etapas principais, descritas a seguir e ilustradas na Figura 1.

1. Seleção das unidades lexicais a serem investigadas;
2. Seleção dos corpora a serem utilizados para o treinamento dos modelos estáticos e ajuste fino dos modelos contextuais;
3. Pré-processamento dos dados;
4. Seleção dos modelos de linguagem e geração dos *embeddings* a partir desses modelos;
5. Análise dos resultados e avaliação da mudança semântica das unidades lexicais ao longo do tempo.

Inicialmente, foram selecionadas unidades lexicais de interesse com a ajuda de um linguista especialista, utilizando critérios que identificam palavras com provável mudança de sentido ao longo do tempo. Em seguida, foi construído um *corpus* diacrônico, chamado Tycholina, a partir da união de dois corpora históricos e contemporâneos, além da construção de um *corpus* a partir de publicações da rede social X (na época Twitter), do Projeto ALiB¹. A aplicação de *embeddings* estáticos e contextuais permitiu a análise da MSL em diferentes janelas de tempo, como será detalhado nas subseções a seguir.

3.1. Seleção das Unidades Lexicais

Para a condução desta pesquisa, foi necessário organizar as lexias investigadas em um grupo limitado de elementos e dividi-las em dois conjuntos, aqui chamados de lexias α e lexias β .

O conjunto de lexias α apresenta unidades lexicais polissêmicas do Atlas Linguístico do Brasil (ALiB), os quais foram segregados a partir da avaliação do especialista linguista e correspondem às palavras com múltiplos sentidos e que podem ser utilizadas em diversos contextos. A seleção cuidadosa desses léxicos permite uma investigação dos diferentes significados atribuídos a eles ao longo do tempo, possibilitando analisar nuances semânticas e dos deslocamentos de sentido que ocorrem na língua.

O Atlas Linguístico do Brasil é um projeto de grandes proporções, idealizado na década de 1950 e desenvolvido em âmbito nacional a partir de 1996. A partir deste projeto foram publicados dois volumes (Cardoso et al., 2014a,b), que descreve a diversidade linguística do Brasil através da cartografia.

Nestes dois volumes são estabelecidas isoglossas, que são fronteiras geográficas que denotam uma certa característica linguística para a delimitação de áreas dialetais. O Volume 2 — *Cartas linguísticas I* (Cardoso et al., 2014b) traz resultados das capitais brasileiras que foram objeto da pesquisa, espelhados em mapas linguísticos, chamados de Cartas, com ricas informações linguísticas representando a realidade variada dos falares do Brasil. A Figura 2 apresenta a Carta BOLINHA DE GUDE do ALiB, em que podem ser observados alguns termos presentes na Carta e os estados brasileiros de sua ocorrência.

A seleção de lexias α foi utilizada na coleta de dados para a construção do *corpus* aqui chamado *ALiB-Twitter*. Em virtude de limitações de coleta, foram selecionados empiricamente unidades lexicais que pudessem contemplar integralmente Cartas do ALiB atendendo ao limite de coleta de publicações da rede social permitidas sobre todo o período disponível. A Tabela 2 mostra os termos utilizados para a coleta, as respectivas Cartas do ALiB a qual pertencem e a quantidade de publicações coletadas da rede social.

No conjunto β as lexias-alvo a serem investigados foram escolhidos com o auxílio de um especialista linguista, segundo os seguintes critérios:

- Termos que hoje designam referentes relativamente recentes, que não existiam em séculos anteriores (incluindo casos que hoje convivem sentidos modernos e sentidos antigos).
- Termos referentes a designações sociais e políticas que podem ter passado por alterações junto com transformações sociais e culturais que ocorreram ao longo dos séculos.

¹ Atlas Linguístico do Brasil

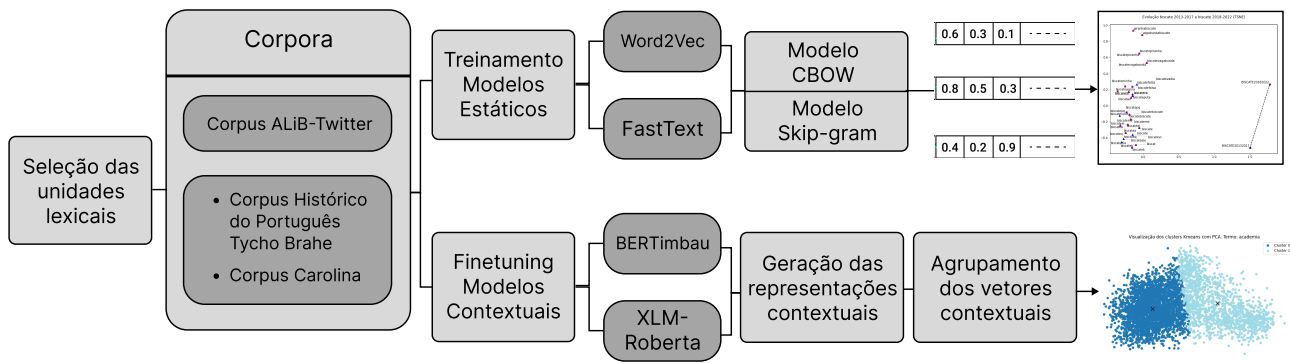


Figura 1: Método de análise da MSL em PT.

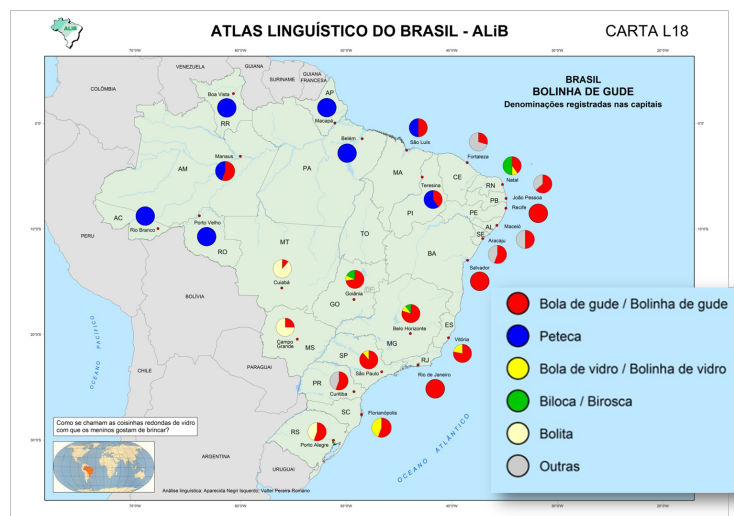


Figura 2: ALiB - Carta BOLINHA DE GUDE (Cardoso & Mota, 2014).

Carta ALiB	Lexia	# publicações coletadas
BICHO DA GOIABA	lagarta	166.674
	broca	515.607
PROSTITUTA	biscate	599.703
CAMBALHOTA	carambola	140.102
	mortal	3.011.949
	perereca	837.362
BOLINHA DE GUDE	peteca	219.714
	birosca	111.888
ESTILINGUE	estilete	99.828
	peteca	*
RUGE	ruge	50.371
Total		5.753.198

* Os mesmos dados foram utilizados para a unidade lexical “peteca” nas cartas “Bolinha de gude” e “Estilingue”.

Tabela 2: Tabela das unidades lexicais utilizadas para a coleta em rede social e o quantitativo de coleta.

- Termos que podem ter passado por mudanças de avaliação social (de positiva para negativa e/ou de negativa para positiva) devido a transformações sociais.

Ao todo foram listadas 31 unidades lexicais polissêmicas possíveis para investigação, contudo, destas destacamos para este trabalho as seguintes: “academia”, “homem”, “indústria”, “memória”, “mestre”, “negro”, “pecado”, “pesoa”, “rede”, “sala”, “virgem”.

3.2. Corpora

Dois *corpora* foram construídos para o desenvolvimento do método:

- *corpus* ALiB-Twitter, composto de publicações da rede social Twitter, atualmente X, contemplando o período dos últimos 10 anos que tivessem a ocorrência das unidades lexicais do conjunto de lexias α .
- *corpus* Tycholina, composto da união dos dados textuais do *corpus* histórico do português Tycho Brahe (Galves, 2018) e do *corpus* Carolina (Crespo et al., 2023).

3.2.1. Corpus ALiB-Twitter

Na construção do *corpus* ALiB-Twitter, a ferramenta Twarc² foi utilizada para realizar a coleta dos dados por meio da API do Twitter. Foram coletados *tweets* publicados entre o período de 01 de janeiro de 2013 a 31 de dezembro de 2022, contemplando uma janela de 10 anos. Além da cuidadosa seleção de palavras-chave e aplicação de filtros para a busca mais precisa, foram excluídas as publicações de publicidade ou patrocinadas e as retransmissões de publicações de outros usuários em perfis pessoais (chamadas *retweets*).

Após a coleta dos dados, uma etapa de limpeza e anonimização dos dados foi necessária, uma vez que, por se tratarem de textos coloquiais de redes sociais, as publicações continham *links*, imagens, *emojis* e caracteres especiais, além de possuírem menções a nomes de usuários. Após esta primeira etapa, os dados foram segmentados para a utilização dos textos no treinamento de modelos de linguagem e geração dos vetores. A Tabela 3 apresenta informações quantitativas dos dados do *corpus* ALiB-Twitter.

Valor total	#
Tweets	5.753.198
Tokens	62.269.473

Tabela 3: Informações do *corpus* ALiB-Twitter.

3.2.2. Corpus Tycholina

O *corpus* Tycholina foi elaborado a partir da união de outros dois *corpora* da língua portuguesa, o *corpus* histórico do português Tycho Brahe (Galves, 2018) e o *corpus* Carolina (Crespo et al., 2023), a fim de expandir o período de tempo investigado, uma vez que o *corpus* ALiB-Twitter cobria uma janela temporal curta e sua construção considerou apenas as unidades lexicais do Projeto ALiB³.

O *corpus* Histórico do Português Tycho Brahe é uma coleção de textos em português escritos por autores nascidos entre 1380 e 1978. Composto por 88 textos (3.544.628 palavras), o *corpus* oferece um rico material para o estudo de mudanças da língua portuguesa, permitindo a análise de mudanças no vocabulário, gramática, sintaxe e expressões idiomáticas. Os textos podem ser acessados via arquivos de extensão *.txt* ou *.xml* e foram editados de acordo com sua natureza.

O *corpus* Carolina, com 823 milhões de *tokens* e textos datados entre 1970 e 2021, complementa o *corpus* Tycho Brahe, expandindo a análise para o período contemporâneo. Sua metodologia, chamada WaC-wiPT, agrega informações detalhadas sobre a origem dos dados coletados e as transformações sofridas no *corpus*, além de classificar os textos em subcorpora por categorias (Sturzeneker et al., 2022). A versão do *corpus* Carolina utilizada foi a v1.2 Ada, que é dividida em sete categorias de dados: *dat* (*datasets* e outros *corpora*), *jud* (jurídico), *leg* (legislativo), *pub* (domínio público), *soc* (mídias sociais), *uni* (domínios de universidade) e *wik* (wikis), destes apenas os ramos *dat*, *pub* e *soc* foram selecionados para integração ao novo *corpus*, por se tratarem de textos mais relacionados ao tipo de texto encontrado no *corpus* Tycho Brahe e conter textos mais próximos ao coloquial. A Tabela 4 apresenta algumas informações quantitativas do *corpus* Tycholina.

Apresentamos a quantidade de dados alocados em cada período para garantir a reprodutibilidade e a transparência em relação à distribuição dos dados ao longo do tempo. O ponto de corte adotado resultou na alocação de 12.376 textos na Janela A e 621.331 textos na Janela B.

²Repositório: <https://github.com/DocNow/twarc>

³<https://alib.ufba.br/>

Essa assimetria reflete a natureza dos dados disponíveis, em que a janela contemporânea (Janela B) apresenta uma concentração expressivamente maior de dados textuais, referentes a interação de domínios da web e redes sociais a partir do *Corpus* Carolina.

Valor total	#
Sentenças	633.801
Tokens	16.221.967

Tabela 4: Informações do *corpus* Tycholina.

3.3. Pré-processamento

Uma vez que os *corpora* foram construídos, o pré-processamento dos dados consistiu na aplicação de técnicas de limpeza, remoção de ruídos, normalização e tokenização, tendo em vista a preparação dos dados para as etapas seguintes considerando que, por conter textos de outras décadas e séculos, há variação também na ortografia. Essa etapa garantiu que os textos estivessem em um formato adequado para a análise a partir dos *embeddings* gerados por meio dos modelos de linguagem selecionados, minimizando a interferência de ruídos nos dados para esta tarefa.

O conjunto de dados do *corpus* ALiB-Twitter cobre um período de 10 anos, já o *corpus* Tycholina tem a cobertura de um período de 641 anos (1380 a 2021). Em ambos os *corpora*, seus respectivos períodos de cobertura foram divididos em duas janelas de tempo, aqui nomeadas **Janelas A e B**. A Tabela 5 apresenta os períodos das divisões de cada janela para os *corpora*. Cada janela temporal é tratada como um *corpus* temporal específico C_A, C_B , partes do *corpus* C completo, de onde serão aprendidas todas as representações vetoriais. Cada lexia-alvo $l \in C$ é substituída por um token temporal específico t_i do tipo l_i , onde $i \in \{A, B\}$. Essa medida torna possível que seja realizado um único treinamento dos modelos estáticos em todo o *corpus*, sem que haja a necessidade de realizar alinhamento de vetores posteriormente, uma vez que, os vetores dos *tokens* em tempos distintos estão representados no mesmo espaço semântico (Dubossarsky et al., 2019).

Para a abordagem contextual foi utilizada uma estratégia distinta, utilizando modelos pré-treinados para português e realizando o ajuste fino destes modelos com o *corpus* Tycholina. Antes do ajuste fino dos modelos contextuais, para cada sentença $s \in C$, foi adicionado o ano de publicação do texto no início de cada sentença.

3.4. Modelos

Duas abordagens computacionais de modelos de linguagem foram utilizadas neste trabalho: modelos estáticos e modelos contextualizados, comumente também chamados de modelos dinâmicos. Os representantes dos modelos estáticos adotados foram o Word2Vec (Mikolov et al., 2013b,c) e Fasttext (Bojanowski et al., 2017b; Joulin et al., 2017b). Estes modelos podem capturar a representação do espaço semântico das unidades lexicais, tendo como base as unidades mais semelhantes em sentido no *corpus* e foram utilizados com os *corpora* ALiB-Twitter e Tycholina. Os modelos contextuais selecionados para os experimentos são baseados na arquitetura BERT, o BERTimbau (Souza et al., 2020) e XLM-Roberta (Conneau et al., 2020). O experimento realizado com esses modelos considerou apenas o *corpus* Tycholina.

Os modelos estáticos foram inicialmente utilizados em função da sua eficiência em gerar representações consistentes, possibilitando uma análise mais concisa e de baixo custo da mudança semântica ao representar cada item lexical como um vetor fixo, calculando a similaridade em diferentes períodos históricos. Entretanto, a complexidade da mudança semântica muitas vezes exige a captura de sutilezas e variações contextuais, tarefa para a qual os modelos contextualizados (ou dinâmicos) são superiores uma vez que podem gerar essas representações dinâmicas do espaço semântico das unidades lexicais com base nas representações contextuais mais semelhantes em sentido no *corpus*. Portanto, este trabalho propõe uma análise comparativa para avaliar os resultados em termos da eficiência dos estáticos e a sensibilidade contextual dos dinâmicos.

Após a seleção das unidades lexicais e da organização dos *corpora*, os modelos estáticos foram treinados; em seguida, o conjunto de termos similares S_l foi analisado. Este conjunto é composto por termos l' cuja similaridade é representada pela função

$$S(t_w, 20) = \{v \in V \mid \text{sim}(v, t_w) \in S_{20}\},$$

em que S_{20} representa os 20 termos com maior similaridade e V é o vocabulário do modelo. Por fim, foram elaboradas visualizações para os *embeddings* em que é possível observar as lexias mais similares as investigadas.

Com relação aos modelos contextualizados, estes foram ajustados individualmente para cada janela de tempo, resultando em dois modelos M_1 e M_2 . Para cada modelo ajustado M_i e para cada lexia-alvo l , o algoritmo *K-means* $\mathbb{K}$ foi aplicado

	<i>corpora</i> ALiB-Twitter	<i>corpora</i> Tycholina
Janela A	01/01/2013 a 31/12/2017	1380 a 1799
Janela B	01/01/2018 a 31/12/2022	1800 a 2021

Tabela 5: Divisão das janelas temporais nos *corpora*.

para criar conjuntos de agrupamentos de sentidos D_l , onde $D_l = \mathbb{K}(M_i, l)$ e as visualizações geradas representam os agrupamentos de sentidos similares considerados pelo algoritmo.

3.5. Protocolo de anotação dos resultados

O processo de anotação e avaliação dos resultados foi realizado de maneira colaborativa por dois anotadores, sendo um deles especialista em linguística (denominado “anotador sênior”) e o outro, “anotador júnior”. Após o treinamento e o *fine-tuning* dos modelos, a caracterização da MSL fundamentou-se na análise das unidades lexicais vizinhas mais próximas a lexia-alvo. O anotador júnior examinava os resultados individualmente e, em seguida, reunia-se semanalmente com o anotador sênior para discutir os casos um a um, buscando consolidar a concordância interanotadores em relação à caracterização da MSL. Embora apenas os casos com concordância plena tenham sido considerados para a análise final, vale ressaltar que não foi mantido um registro quantitativo ou uma listagem exaustiva das divergências iniciais de interpretação. Essa ausência decorre da própria dinâmica qualitativa dos encontros interanotadores, onde as deliberações frequentemente englobavam abstrações teóricas e o levantamento de hipóteses que, mesmo documentadas de forma exploratória, não chegaram a ser testadas empiricamente com os modelos computacionais. Sendo assim, o protocolo de anotação priorizou a construção do consenso para as unidades lexicais analisadas, não sendo viável ou metodologicamente previsto quantificar as instâncias de não concordância ocorridas.

4. Experimentos

O design experimental foi realizado considerando duas abordagens metodológicas distintas. A primeira abordagem concentrou-se no uso de modelos semânticos distribucionais estáticos. Modelos que, baseados na hipótese da coocorrência dos *tokens*, geram *embeddings* estáveis capazes de capturar relações semânticas e sintáticas globais no vocabulário do *corpus*. A aplicação desses modelos a *corpora* diacrônicos permite o rastreamento

das transformações de significado das lexias ao longo de períodos históricos definidos. A segunda abordagem utilizou o paradigma dos *embeddings* contextuais dinâmicos. Foram explorados modelos de linguagem baseados na arquitetura BERT já pré-treinados para a língua portuguesa. Essa escolha metodológica se beneficia da capacidade inerente desses modelos de gerar representações sensíveis ao contexto, possibilitando uma análise mais granular e abrangente. Através destes dois experimentos, pretende-se responder as duas questões de pesquisa, já mencionadas:

1. Qual é a eficiência dos modelos estáticos em identificar mudanças semânticas (*semantic shift*)?
2. De que forma os modelos contextuais permitem caracterizar a mudança semântica (*semantic drift*)?

A partir destas questões de pesquisa, o presente estudo se desdobrou em dois experimentos principais:

- **Experimento 1:** Com o *corpus* ALiB-Twitter foram usados exclusivamente modelos estáticos para avaliar a vitalidade (ou seja, se as lexias mantém o uso no sentido catalogado no ALiB) e a transformação semântica do conjunto de lexias α (*semantic shift*).
- **Experimento 2:** Com o *corpus* Tycholina, empregou-se tanto modelos estáticos como contextuais em um estudo comparativo, dedicado à caracterização da mudança semântica no conjunto de lexias β (*semantic drift*).

É importante salientar uma limitação do design experimental desta pesquisa, referente à assimetria na aplicação dos modelos aos *corpora*. Inicialmente, o escopo previa a análise exclusiva do conjunto de lexias α . Contudo, após a construção do *corpus* ALiB-Twitter, que teve sua coleta estritamente direcionada aos itens específicos do conjunto, o acesso à API da rede social, de onde os dados eram coletados, foi revogado, impossibilitando a extração de novos dados para abranger o conjunto de lexias β . Dada a baixa frequência dos termos do conjunto β no *corpus* ALiB-Twitter, foi adotada a elaboração do *corpus* Tycholina para viabilizar a continuidade do

estudo, por fornecer um volume maior de dados e cobertura temporal. Essa variação pode limitar uma comparação experimental controlada. Consequentemente, as diferenças de desempenho observadas, devem ser interpretadas considerando não apenas a estabilidade textual dos documentos, mas também variáveis como tamanho dos *corpora* e do vocabulário.

4.1. Treinamento e Parametrização dos Modelos de *Embeddings* Estáticos

O treinamento dos modelos semânticos distribucionais estáticos foi conduzido de maneira independente para cada conjunto de dados, utilizando os *corpora* ALiB-Twitter e Tycholina. Essa abordagem resultou em um conjunto de modelos individualizados por *corpus* e período temporal específico, permitindo a análise comparativa de diferentes janelas históricas.

A implementação para geração dos *embeddings* foi realizada utilizando a biblioteca de código aberto Gensim 4.3.2 (Řehůřek & Sojka, 2010). Para os modelos Word2Vec e FastText, foram exploradas ambas as arquiteturas de treinamento: Continuous Bag-of-Words (CBOW) e Skip-gram.

A parametrização dos modelos seguiu melhores práticas e recomendações da literatura e foi ajustada como segue:

1. Dimensionalidade do vetor (d):

Fixada em 100 dimensões.

2. Janela de contexto (*window size*):

Foi adotada uma janela simétrica de 5 unidades lexicais (± 5), seguindo a observação de Levy & Goldberg (2014), para capturar de forma mais ampla as relações semânticas.

3. Épocas de treinamento (*epochs*):

3 épocas.

4. Frequência mínima (*min count*):

Foi definido um corte de frequência mínima de 20 ocorrências. Termos com frequência inferior a esse critério foram ignorados, visando mitigar introdução de ruídos por *tokens* esparsos, que poderiam comprometer a qualidade das representações geradas.

A seleção de parâmetros adotada buscou equilíbrio entre granularidade da representação e a qualidade estatística dos espaços vetoriais gerados. Embora haja, na literatura, vários trabalhos que referenciem a dimensionalidade $d = 300$ para modelos treinados em grandes bases de dados, a adoção de $d = 100$ é adequada para os *corpora* utilizados no presente trabalho. Essa escolha

buscou viabilizar representações consistentes, capazes de capturar os deslocamentos semânticos. Ademais, o limite $min_count = 20$ foi adotado com o objetivo de garantir a estabilidade das representações, filtrando flutuações e possíveis erros ortográficos, apesar de este ser uma possível inserção de viés de frequência que pode limitar a sensibilidade dos modelos em identificar a MSL. Dessa forma, foi priorizada a confiabilidade metodológica das vizinhanças semânticas globais em detrimento da captura de *shifts* isolados em *tokens* de baixa frequência.

O processo de treinamento dos modelos estáticos resultou em 4 modelos de *embeddings* treinados e parametrizados uniformemente. Esses modelos estabelecem a base vetorial *type-level* do estudo, porém, é necessário notar que essa natureza estática das representações limita a captura de variação de sentido em função do contexto imediato de uso.

4.2. Ajuste Fino, Parametrização e Extração de *Embeddings* Contextuais Dinâmicos

O segundo experimento se concentrou na utilização de modelos capazes de gerar representações contextuais/dinâmicas. Os modelos selecionados para essa etapa foram o BERTimbau (Souza et al., 2020) e o XLM-RoBERTa (Conneau et al., 2020), sendo BERTimbau um modelo pré-treinado para o português e o outro multilíngue.

O *corpus* Tycholina foi pré-processado para garantir que as instâncias fossem processadas dentro das limitações de sequência dos modelos BERT. Nessa etapa foi realizada a filtragem das sentenças que continham as lexias-alvo. Para geração dos *embeddings*, a janela de contexto foi padronizada, tendo a lexia-alvo centralizada e a sentença truncada em 64 *tokens* de contexto à esquerda e à direita, totalizando ≤ 129 *tokens*, incluindo o alvo. O objetivo dessa estratégia foi minimizar o ruído e garantir a relevância do contexto na janela de atenção dos modelos.

Antes da extração dos vetores, foi realizado um ajuste fino dos modelos BERT e XLM-RoBERTa utilizando o *corpus* Tycholina. A parametrização para o *fine-tuning* foi ajustada segundo o seguinte:

1. Tarefa de ajuste fino:

Foi utilizada a tarefa de *Masked Language Modeling* (MLM), para adaptação dos modelos ao domínio diacrônico dos dados.

2. Taxa de aprendizado (*learning rate*):

Definida em $2e^{-5}$.

3. Tamanho do lote (*batch size*):

8 instâncias por lote.

4. Épocas de treinamento (*epochs*):

3 épocas.

Após o ajuste fino, os vetores contextuais para as ocorrências das lexias-alvo foram extraídas da última camada do modelo. Por fim, esses *embeddings* foram submetidos ao algoritmo de agrupamento *K-means* para alocar as instâncias a conjuntos de sentidos similares, permitindo a visualização bidimensional da dispersão no espaço semântico.

É importante ressaltar como a comparabilidade temporal é garantida metodologicamente nas duas abordagens adotadas. Para os modelos estáticos, a adoção da técnica de Referência Temporal (Dubossarsky et al., 2019) assegura que os vetores aprendidos estejam em um espaço semântico compartilhado, eliminando a necessidade de alinhamento dos vetores em etapas posteriores.

No caso dos modelos contextuais, a comparabilidade não é feita pela sobreposição direta das coordenadas vetoriais, mas pela análise da topologia dos sentidos por meio de agrupamentos (Martinc et al., 2020). A comparação entre as janelas temporais é baseada na análise da distribuição dos vizinhos semânticos para cada lexia-alvo, sendo uma alternativa para contornar a incomparabilidade geométrica em relação aos modelos ajustados em épocas distintas (Kutuzov & Giulianelli, 2020). Sendo assim, a caracterização tipológica da mudança é baseada em uma validação qualitativa das relações semânticas de cada período, em vez da aplicação de métricas de distância geométrica entre valores de modelos independentes.

4.3. Representação gráfica dos espaços vetoriais

Para a visualização do deslocamento semântico das unidades lexicais, foram utilizadas técnicas de redução de dimensionalidade aplicadas aos *embeddings* gerados pelos modelos. Especificamente, o t-SNE (*t-distributed Stochastic Neighbor Embedding*) foi aplicado para mapear o espaço vetorial de alta dimensionalidade dos modelos em um plano bidimensional. Cada lexia é representada por um ponto neste espaço semântico. Essa projeção foi utilizada como apoio para a avaliação qualitativa da proximidade e dispersão vetorial, complementando a métrica.

Com o intuito de garantir a legibilidade visual das projeções e evitar sobreposição de rótulos, foi adotada uma estratégia de otimização das representações. Nas figuras apresentadas ao longo das seções de resultados, especialmente no que diz respeito aos modelos estáticos, foi adotado um critério de filtragem para exibição de *labels*, destacando apenas os vizinhos mais próximos a lexia-alvo.

5. Resultados

Esta seção apresenta os resultados obtidos por meio dos experimentos 1 e 2. A análise dos resultados estabelece uma distinção entre os *corpora*: *corpus* ALiB-Twitter e *corpus* Tycholina. As subseções subsequentes apresentam os resultados e uma discussão mais detalhada na identificação e categorização da MSL em cada um dos experimentos citados na Seção 4.

5.1. Experimento 1: *corpus* ALiB-Twitter

A primeira etapa da análise dos resultados teve como objetivo a avaliação da vitalidade e transformação semântica dos termos do ALiB (Lexias α) na rede social Twitter (atualmente X), a partir do *corpus* ALiB-Twitter, utilizando modelos semânticos distribucionais estáticos treinados com Word2vec e FastText. O rastreo diacrônico foi realizado por meio da Similaridade de Cosseno aplicada aos vetores das unidades lexicais do conjunto α entre as Janelas A e B. Os resultados obtidos por cada modelo estão detalhados nas Tabelas 6 e 7, respectivamente.

Uma primeira observação a ser discutida aponta para a divergência dos modelos na identificação da mudança (ex: "broca", "birosca") - *semantic shift*. O modelo Word2Vec detectou um único caso de MSL, o da lexia "biscate", caracterizado quanto à mudança como ampliação em direção a sentidos de parentesco. Por outro lado, o modelo FastText não identificou a MSL das lexias, mas apresentou maior sensibilidade à vitalidade de "broca".

A análise das Tabelas 6 e 7 expõe que a detecção de vitalidade é sensível à arquitetura do modelo de embedding. O modelo FastText demonstrou um melhor desempenho na atribuição de vitalidade, o que pode ser atribuído à sua capacidade de incorporar informações de *subword* nas representações geradas. Em *corpora* de redes sociais como o ALiB-Twitter, que frequentemente contêm ruído, erros ortográficos e variação morfológica, a modelagem baseada em morfemas do FastText tende a gerar representações mais

estáveis para *tokens* de baixa frequência, inexistentes ou variantes lexicais do vocabulário do modelo, mitigando o problema da dispersão de dados que afeta o Word2Vec em ambientes ruidosos.

A identificação de MSL no *corpus* ALiB-Twitter é esparsa, com apenas a unidade lexical “biscate” classificado como “ampliação” pelo modelo Word2Vec. A observação de que esta mudança se dá em direção a sentidos de graus de parentesco sugere uma evolução semântica complexa, característica de contextos dialetais ou sociolinguísticos do português. No entanto, a falha do FastText em replicar essa detecção levanta questões. Embora o FastText tenha tido um desempenho melhor na identificação da vitalidade, sua média vetorial pode ser suavizada pela informação de subword, apagando mudanças sutis de contexto global.

As divergências observadas reforçam que a escolha da arquitetura de embedding deve ser calibrada em estudos diacrônicos que utilizam dados de mídias sociais. Em um contexto no qual a MSL está ligada à vitalidade da unidade lexical, o FastText pode oferecer uma base para identificar quais termos merecem análise. Contudo, o Word2Vec demonstrou maior sensibilidade a uma mudança semântica específica.

Parte da análise foi realizada a partir das representações gráficas dos espaços semânticos dos conjuntos de lexias analisadas. Essas projeções podem ser visualizadas nas Figuras 3 e 4, em relação as unidades lexicais destacadas “biscate” (Word2Vec) e “broca” (FastText), respectivamente.

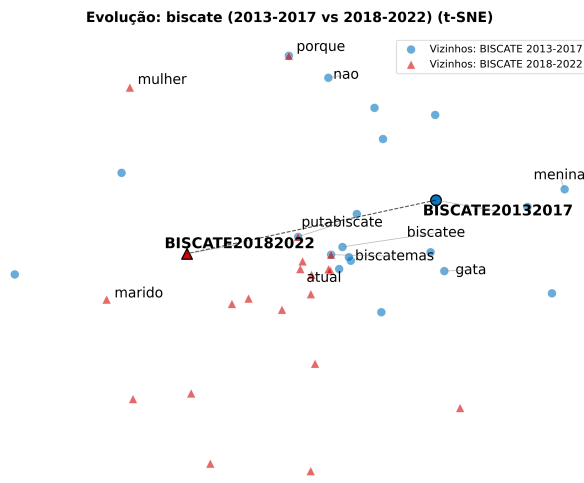


Figura 3: Representação da unidade lexical “biscate” com modelo Word2Vec.

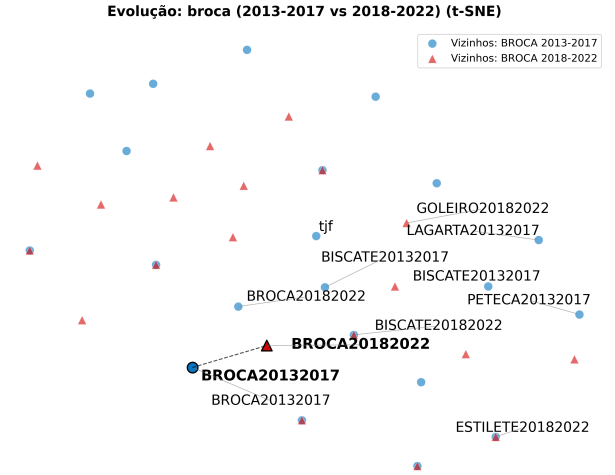


Figura 4: Representação da unidade lexical “broca” com modelo Fasttext.

5.2. Experimento 2: *Corpus* Tycholina

A segunda fase da análise de resultados concentra-se na investigação da MSL em um *corpus* histórico, que abarca uma maior janela de tempo e quantidade de dados (*corpus* Tycholina). Esta etapa visa confrontar a aplicação e a resolução das abordagens estática e contextual. Serão apresentados os achados para os conjuntos de lexias α e β em relação ao *corpus* Tycholina. Enquanto os modelos estáticos (Word2Vec e FastText) continuam a oferecer a quantificação global da MSL por meio da Similaridade de Cosseno, a aplicação dos *embeddings* contextuais permite uma análise utilizando um algoritmo de agrupamento (*K-means*) e visualização bidimensional.

5.2.1. Método Estático

Os resultados obtidos no Experimento 2, a partir dos *embeddings* estáticos, com o conjunto de Lexias α , contrastam com as observações do Experimento 1 visto que a análise mostra uma estabilidade semântica. Conforme apresentado nas Tabelas 8 e 9, o *corpus* Tycholina não apresentou nenhum caso de MSL detectável por qualquer uma das arquiteturas estáticas. Embora essa ausência de *shift* semântico corrobore a hipótese de que *corpora* com textos mais estáveis e formais demonstram uma dinâmica lexical mais lenta em comparação com ambientes voláteis em tendências, como os das mídias sociais (*corpus* ALiB-Twitter), é necessário ressaltar as limitações do experimento, especialmente no que se refere aos resultados não refletirem exclusivamente a natureza do gênero textual, sendo assim, os resultados obtidos podem também ser consequência do vocabulário, volume dos dados ou

Modelo Word2Vec - Léxias α - <i>corpus</i> ALiB-Twitter				
Lexia	Vitalidade	Mudança	Tipo	Observação
lagarta	✓	×	—	—
broca	×	×	—	—
biscate	✓	✓	ampliação	Em direção a sentidos de graus de parentesco
carambola	×	×	—	Similares no sentido de fruta
mortal	×	×	—	—
perereca	×	×	—	—
peteca	✓	×	—	—
birosca	✓	×	—	—
estilete	×	×	—	—
ruge	×	×	—	—

Tabela 6: Resultados obtidos com Word2Vec com as lexias α no *corpus* ALiB-Twitter

Modelo FastText - Lexias α - <i>corpus</i> ALiB-Twitter				
Lexia	Vitalidade	Mudança	Tipo	Observação
lagarta	✓	×	—	—
broca	✓	×	—	—
biscate	✓	×	—	—
carambola	×	×	—	—
mortal	×	×	—	—
perereca	×	×	—	—
peteca	✓	×	—	—
birosca	×	×	—	—
estilete	×	×	—	—
ruge	×	×	—	—

Tabela 7: Resultados obtidos com FastText com lexias α no *corpus* ALiB-Twitter.

da frequência de ocorrência das unidades lexicais nos textos históricos do *corpus* Tycholina.

Entretanto, a avaliação da vitalidade das unidades lexicais ainda expõe diferenças entre os modelos. O FastText (Tabela 9), se mostrou eficaz em identificar a vitalidade em cinco lexias (“lagarta”, “broca”, “carambola”, “peteca” e “birosca”), ao contrário do modelo Word2Vec, que não detectou. Os resultados obtidos para a unidade lexical de “carambola” reforçam que a modelagem de *subword*, presente no modelo FastText, indica uma maior resiliência na representação das unidades lexicais, uma vez que, em *corpora* onde os termos podem aparecer em contextos esparsos ou especializados, a composição de n-gramas permite que o modelo gere representações, mesmo quando há baixa frequência da lexia, contornando a limitação de palavras fora do vocabulário de treino e capturando informações de associação semântica.

Além dos experimentos com o conjunto de Lexias α , foram realizadas análises com o conjunto de Lexias β no mesmo *corpus*; entretanto, para esse grupo de unidades lexicais, não foi avaliada

a vitalidade semântica. Diferente dos resultados observados com o conjunto de Lexias α , o modelo Word2Vec (Tabela 10) demonstrou uma alta taxa de detecção da mudança (*shift semântico*) em 8 de 11 unidades lexicais investigadas. O FastText (Tabela 11) apresentou uma menor sensibilidade, detectando essa MSL em 6 de 11 unidades lexicais investigadas.

Para além do quantitativo de lexias das quais foram identificadas essa MSL, é importante destacar que os modelos, especialmente o Word2Vec, conseguiram capturar transformações com implicações sociais e tecnológicas como a ampliação do sentido de “memória” e de “rede” refletindo a inclusão digital tratada nos textos; a lexia “homem” que sofreu uma pejoração; a unidade lexical “academia” que se expandiu do sentido de formação educacional, enquanto “mestre” sofreu uma bifurcação em sentido de “grau acadêmico”. As diferenças de detecção entre os modelos sugere que a escolha do modelo deve ser guiada pela natureza esperada da MSL em análise.

Parte da análise apresentada nesta seção foi realizada a partir das representações gráficas dos

Modelo Word2Vec - Lexias α - <i>corpus</i> Tycholina				
Lexia	Vitalidade	Mudança	Tipo	Observação
lagarta	×	×	—	A lexia foi encontrada em ambas as janelas A e B, porém com o sentido de fatalidade e não de camalhota como previsto no ALIB.
broca	×	×	—	
biscate	×	×	—	
carambola	×	×	—	
mortal	×	×	—	
perereca	×	×	—	
peteca	×	×	—	
birosca	×	×	—	
estilete	×	×	—	
ruge	×	×	—	

Tabela 8: Resultados obtidos com Word2Vec com lexias α no *corpus* Tycholina.

Modelo FastText - Lexias α - <i>corpus</i> Tycholina				
Lexia	Vitalidade	Mudança	Tipo	Observação
lagarta	✓	×	—	Unidades lexicais com sentido de alimento (“sorvete”, “bacalhau”)
broca	✓	×	—	
biscate	×	×	—	Unidades lexicais que iniciam com “ca” ou “car”
carambola	✓	×	—	
mortal	×	×	—	
perereca	×	×	—	
peteca	✓	×	—	
birosca	✓	×	—	
estilete	×	×	—	
ruge	×	×	—	

Tabela 9: Resultados obtidos com FastText com lexias α no *corpus* Tycholina.

espaços semânticos dos conjuntos de lexias analisadas. Essas projeções, cujos resultados foram discutidos anteriormente, podem ser visualizadas nas Figuras 5 a 14, em relação as unidades lexicais destacadas nesta subseção.

5.2.2. Método Contextual

O uso dos métodos contextuais com o conjunto de Lexias α apresenta uma concordância entre os modelos utilizados. As Tabelas 12 (BERTimbau) e 13 (XLM-RoBERTa) mostram que os modelos detectaram vitalidade e mudança apenas na unidade lexical “mortal”. Esse resultado sugere que os modelos contextuais podem ter capturado um desvio no uso contextual dessa unidade lexical em sentenças ao longo do tempo. Essa capacidade reforça a sensibilidade desses modelos em identificar pequenas alterações de uso que a visão agregada dos embeddings estáticos pode ignorar.

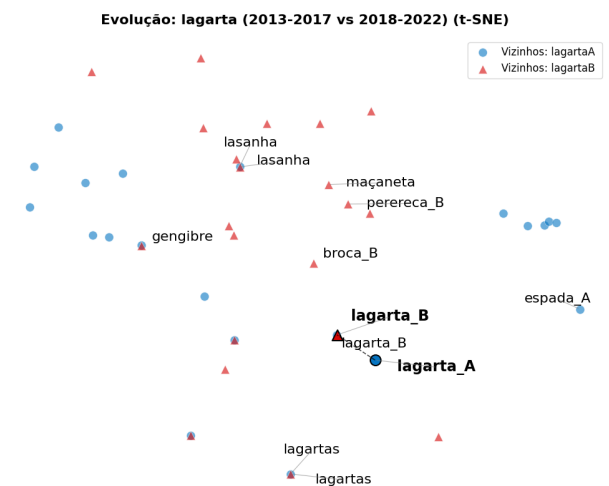


Figura 5: Representação da unidade lexical “lagarta” com modelo FastText.

Modelo Word2Vec - Lexias β - <i>corpus</i> Tycholina			
Lexia	Mudança	Tipo	Observação
academia	✓	ampliação	De “formação educacional” a “musculação”, “ginástica”, “pilates”
homem	✓	pejoração	Na Janela A, conotações neutras ou positivas, na Janela B pejorativas como “agressor”, “bandido”, “bêbado”
indústria	✓	mudança	De “posição política” para “industrialização” e “empresas”
memória	✓	ampliação	Do sentido de “lembrança” para o sentido de “armazenamento”, “aparelho”
mestre	✓	bifurcação	De uma titulação social para “grau acadêmico”
negro	✓	ampliação	Personagens fictícios e aspectos históricos e culturas na Janela B
pecado	×	–	Ganha uma conotação mais moral, com uma maior presença de termos relacionados à violação de preceitos e desobediência.
pessoa	×	–	–
rede	✓	ampliação	Na Janela B, associação forte com comunicação, TV, redes sociais.
sala	×	–	Na Janela B há uma especialização no sentido de “cômodo”
virgem	✓	mudança	Do sentido de “casta” para o sentido de “signo do zodíaco”

Tabela 10: Resultados obtidos com Word2Vec com lexias β no *corpus* Tycholina.

Modelo FastText - Léxias β - <i>corpus</i> Tycholina			
Lexia	Mudança	Tipo	Observação
academia	×	–	“Faculdade”, “mestre”, “escola” aparecem em A (CBOW) e “ginástica” aparece em B (Skip-gram)
homem	×	–	Presença de nomes de personagens fictícios na Janela B.
indústria	×	–	Processo de industrialização e empresas
memória	✓	ampliação	Do sentido de “lembrança” para o sentido de “armazenamento”, “aparelho”
mestre	✓	ampliação	Do sentido de “erudito” para “grau acadêmico”
negro	×	–	Na Janela B, presença de unidades lexicais como “pobre”, “racista”, “herói”.
pecado	✓	ampliação	De questões religiosas (“cristo”, “deus”) para questões psicológicas (“paranóia”, “ciúme”).
pessoa	×	–	Muitos termos lexicalmente similares a “pessoa”
rede	✓	mudança	Na Janela B, “internet”, “reddit”, “linkedin” e “redeglobo”, “redetv”
sala	✓	ampliação	Na Janela B há uma especialização no sentido de “cômodo”
virgem	✓	mudança	Do sentido de “casta” para o sentido de “horóscopo”

Tabela 11: Resultados obtidos com FastText com lexias β no *corpus* Tycholina.

A avaliação da MSL do conjunto de Lexias β no *corpus* Tycholina, com os modelos contextuais, confirma muitas das transformações semânticas detectadas anteriormente. As Tabelas 14 (BERTimbau) e 15 (XLM-RoBERTa) demonstram uma concordância entre os modelos, os quais identificaram seis (“academia”, “homem”, “memória”, “mestre”, “rede”, “virgem”) e oito (“indústria” e “negro”) casos, respectivamente.

Os resultados também sugerem uma vantagem do método em sua capacidade de gerar *embeddings* dinâmicos para cada instância de uso, em contraste com os métodos estáticos que calculam uma média global para cada janela temporal. O uso do agrupamento possibilita que as instâncias com novos sentidos (e.g., “memória” com sentido de “componente eletrônico” em T_2) possam ser separadas no espaço semântico do sentido prévio. As observações em “negro” (“Rio

Método BERTimbau - Léxias α - <i>corpus</i> Tycholina				
Lexia	Vitalidade	Mudança	Tipo	Observação
lagarta	×	×	—	—
broca	×	×	—	—
biscate	×	×	—	—
carambola	×	×	—	Lexias com sentido de frutas.
mortal	✓	✓	—	Lexias de conotação negativa.
perereca	×	×	—	—
peteca	×	×	—	—
birosca	×	×	—	—
estilete	×	×	—	—
ruge	×	×	—	—

Tabela 12: Resultados do Modelo BERTimbau com Lexias α no *corpus* Tycholina.

Método XLM-RoBERTa - Lexias α - <i>corpus</i> Tycholina				
Lexia	Vitalidade	Mudança	Tipo	Observação
lagarta		×	×	—
broca		×	×	—
biscate		×	×	—
carambola		×	×	Lexias com sentido de alimentos.
mortal		✓	✓	Lexias de conotação negativa.
perereca		×	×	—
peteca		×	×	—
birosca		×	×	—
estilete		×	×	—
ruge		×	×	—

Tabela 13: Resultados do Modelo XLM-RoBERTa com Lexias α no *corpus* Tycholina.

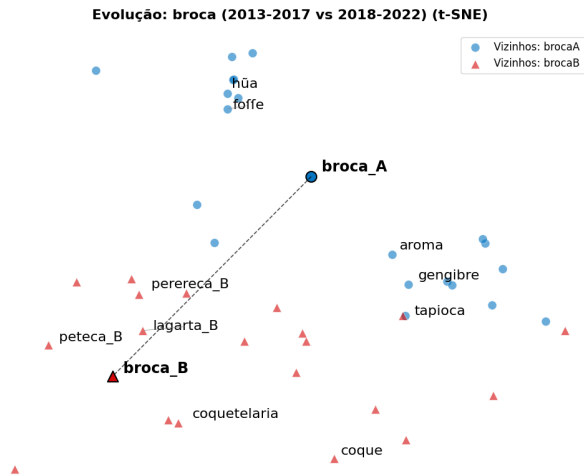


Figura 6: Representação da unidade lexical “broca” com modelo FastText.

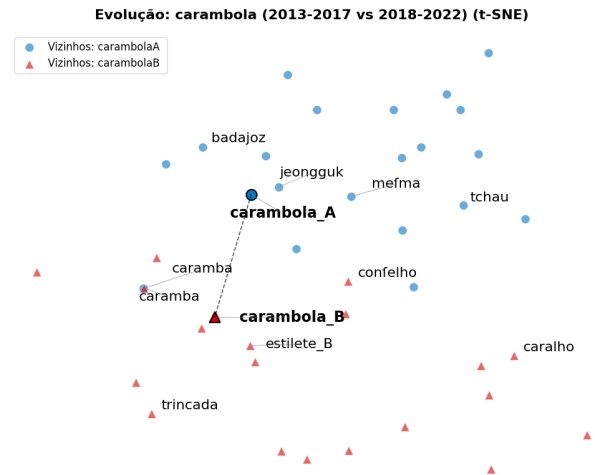


Figura 7: Representação da unidade lexical “carambola” com modelo FastText.

Negro”, “buraco negro”) e “pecado” (uso figurativo) demonstram que os modelos podem representar a pluralidade de sentidos de uma determinada unidade lexical, sem necessariamente classificar como MSL, aprofundando a análise de uso e contexto além do deslocamento vetorial.

Modelo BERTimbau - Léxias β - <i>corpus</i> Tycholina			
Lexia	Mudança	Tipo	Observação
academia	✓	ampliação	Sentido de instituição de ensino T1 para inclusão de sentido de lugar de prática de atividade física
homem	✓	pejoração	Em T2 um dos grupos possui sentido pejorativo relacionados a violência
industria	×	–	Processo de industrialização e empresas
memoria	✓	ampliação	Em T2 surge o sentido relacionado a “componente eletrônico”
mestre	✓	ampliação	Do sentido de “erudito” para “titulação acadêmica”
negro	×	–	Em T2 há presença de “Rio Negro” e “buraco negro”
pecado	×	–	Em T2 temos uso figurativo do item lexical
pessoa	×	–	–
rede	✓	bifurcação	Em T1 há o estreitamento de “conjunto de pontos interligados” para “rede social”
sala	×	–	–
virgem	✓	ampliação	Além do sentido “pureza”, T2 apresenta o sentido de horóscopo

Tabela 14: Resultados do Modelo BERTimbau com Lexias β no *corpus* Tycholina.

Modelo XLM-RoBERTa - Léxias β - <i>corpus</i> Tycholina			
Lexia	Mudança	Tipo	Observação
academia	✓	ampliação	De “formação educacional” para o sentido de lugar de prática de atividades físicas
homem	✓	pejoração	Na Janela A, conotações neutras ou positivas, na Janela B pejorativas como “agressor”, “bandido”, “bêbado”
industria	✓	mudança	De “posição política” para “industrialização”
memoria	✓	ampliação	Do sentido de “lembrança” para o sentido de “armazenamento”, “aparelho” e outros relacionados a tecnologia
mestre	✓	bifurcação	De um título social para “grau acadêmico”
negro	✓	ampliação	Personagens fictícios e aspectos históricos e culturas na Janela B
pecado	×	–	Ganha uma conotação mais moral, com uma maior presença de termos relacionados à violação de preceitos e desobediência.
pessoa	×	–	–
rede	✓	ampliação	Na Janela B, associação forte com comunicação, TV, redes sociais.
sala	×	–	Na Janela B há uma especialização no sentido de um cômodo residencial
virgem	✓	mudança	Do sentido de “casta” para o sentido de “signo do zodíaco”

Tabela 15: Resultados do Modelo XLM-RoBERTa com Lexias β no *corpus* Tycholina.

6. Discussão

A identificação da Mudança Semântica Lexical (MSL) através dos modelos estáticos possui desafios inerentes à arquitetura destes modelos, que geram representações vetoriais fixas e agregadas, impossibilitando a modelagem explícita da variação de sentido ou da polissemia que uma unidade lexical pode apresentar simultaneamente em diferentes janelas temporais. Os resultados apresentados, especialmente a baixa detecção de

MSL no conjunto α do *corpus* Tycholina, indicam que a análise de coocorrência, neste caso, pode suavizar mudanças semânticas ou detectar mudanças que não refletem a variação de uso, mas sim ruído na vizinhança lexical, conforme observado na divergência entre Word2Vec e FastText (Tabelas 8 e 9). Além disso, a análise oferece uma visão média, exigindo uma validação qualitativa exaustiva dos resultados para inferir a real alteração de sentido.

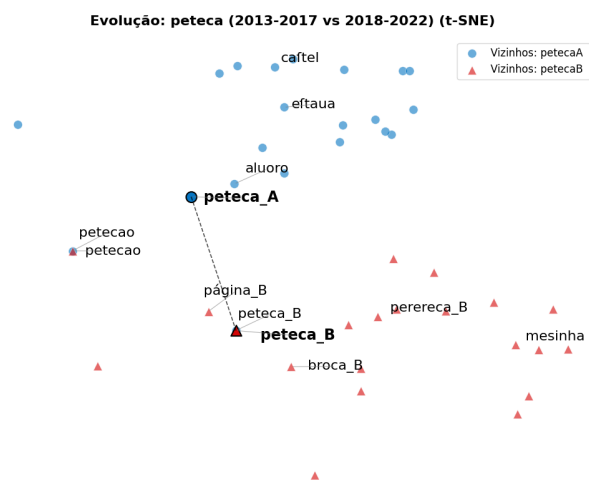


Figura 8: Representação da unidade lexical “peteca” com modelo FastText.

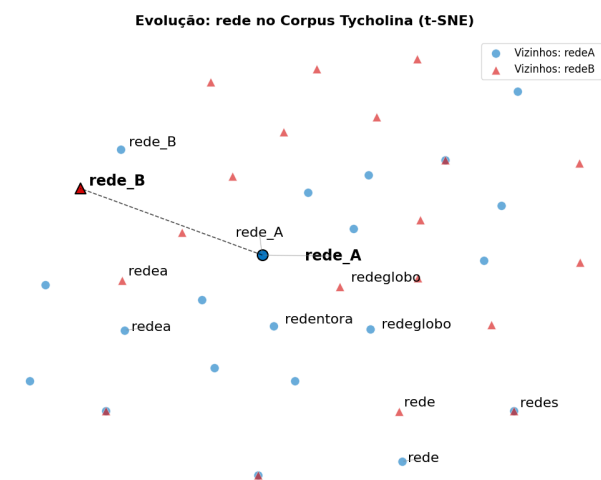


Figura 11: Representação da unidade lexical “rede” com modelo FastText.

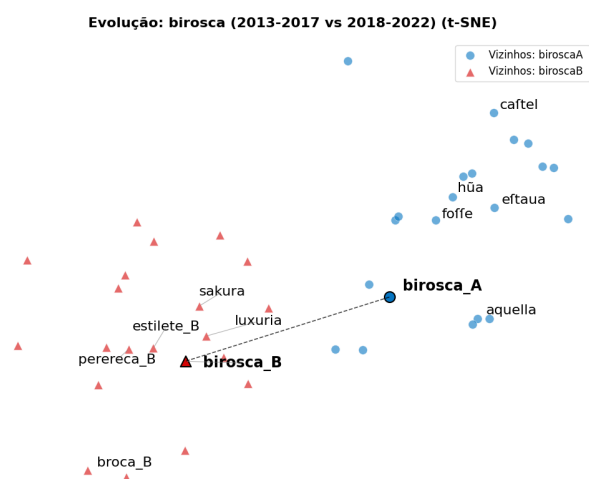


Figura 9: Representação da unidade lexical “birosca” com modelo FastText.

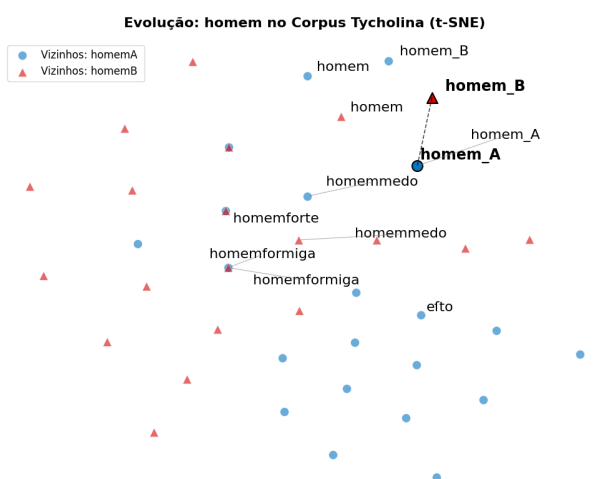


Figura 12: Representação da unidade lexical “homem” com modelo FastText.

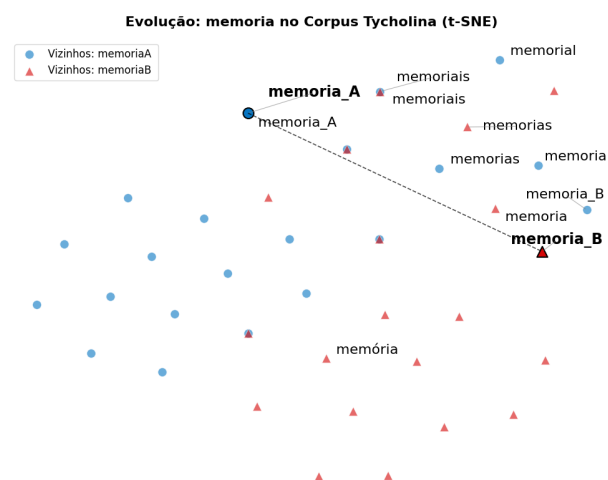


Figura 10: Representação da unidade lexical “memória” com modelo FastText.

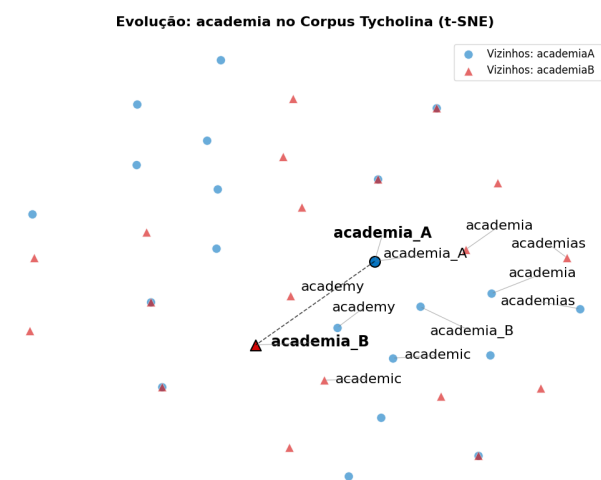


Figura 13: Representação da unidade lexical “academia” com modelo FastText.

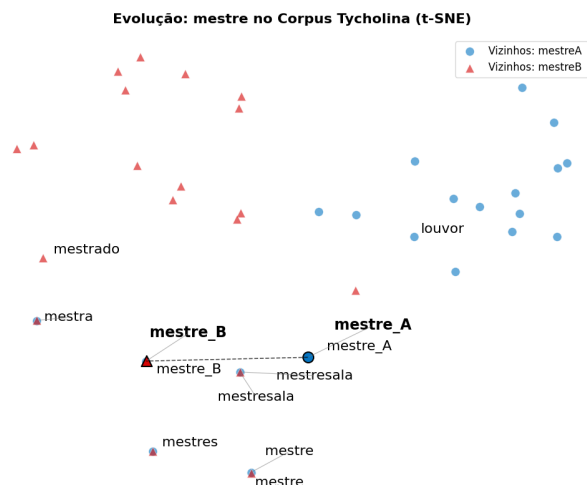


Figura 14: Representação da unidade lexical “mestre” com modelo FastText.

A ausência de detecção de MSL na lexia “pessoa” por todos os quatro modelos testados está de acordo com a da Lei da Conformidade Hamilton et al. (2016b). Por se tratar de uma unidade lexical de alta frequência, sua taxa de mudança semântica é inversamente proporcional à sua recorrência, resultando na não-detecção de mudança por todo o período investigado.

Entretanto, as lexias “pecado” e “sala” apresentam um cenário de detecção isolada, em que apenas o modelo FastText sinalizou o fenômeno de ampliação. Essa questão não invalida a análise, mas demonstra as particularidades arquiteturais dos modelos de espaço semântico. A modelagem baseada na composição de *n-gramas* favorece que o FastText tenha uma sensibilidade maior para capturar correlações periféricas em contextos esparsos, contornando ruídos e variações ortográficas comuns em textos históricos. A falta de consenso com o Word2Vec e com os modelos contextuais dinâmicos sugere que a mudança capturada não representa uma reestruturação profunda ou generalizada do espaço semântico, mas sim variações distribucionais sutis.

Em contraste com o que foi observado nos modelos estáticos, os modelos contextuais superaram essa limitação de representações estatísticas ao produzir embeddings distintos para o mesmo item lexical, refletindo o uso desse item em sentenças específicas, conforme o contexto. O uso de técnicas de agrupamento nesses embeddings permite o mapeamento de grupos vetoriais que correspondem aos diferentes sentidos que uma unidade lexical pode ter em uma janela de tempo. As alterações nesses grupos ao longo do tempo sugerem as transformações semânticas dos sentidos no espaço semântico. Os resultados apre-

sentados (Tabelas 14 e 15) demonstram essa premissa; porém, a interpretação desses conjuntos de embeddings continua sendo uma tarefa complexa que, assim como nos modelos estáticos, necessita de uma avaliação qualitativa para garantir a validade linguística dos agrupamentos gerados.

Outra questão a ser observada no *pipeline* de análise dos modelos contextuais está nas limitações do próprio algoritmo de agrupamento empregado. Foi utilizado o algoritmo *K-means*, cuja eficácia depende intrinsecamente da correta parametrização do número de grupos (k). No presente trabalho, foi utilizado como parâmetro k o valor correspondente do Coeficiente de Silhueta, os quais podem ser observados na Tabela 16. Embora o Coeficiente de Silhueta tenha sido utilizado como guia heurístico para a seleção do parâmetro k , o *K-means* impõe a restrição de $k \geq 2$. Essa imposição algorítmica introduz o risco de sobre-segmentar o espaço semântico, fazendo com que um único grupo semântico seja artificialmente dividido.

Ao observar os valores da Tabela 16 é possível capturar algumas evidências empíricas e comportamentos das arquiteturas utilizadas. É possível observar a predominância do modelo XLM-RoBERTa para o limite inferior ($k = 2$) em quase todas as unidades lexicais e janelas temporais. Esse padrão pode ser um indicativo de que, na ausência de mudanças semânticas drásticas, o algoritmo agrupou espaços com alta coesão estrutural. Em contraste, o modelo BERTimbau apresentou dispersão maior com segmentações de alta granularidade. Essa discrepância pode ser um indicativo de que o modelo monolíngue possui maior sensibilidade às variações contextuais, que são refletidas no espaço semântico. Entretanto, essa sensibilidade pode aumentar o risco de sobre-segmentação.

Além disso, é importante discutir o impacto da assimetria temporal do *corpus* Tycholina sobre os resultados. O desequilíbrio entre as janelas pode introduzir ruídos nos modelos. A maior variação de domínios na Janela B tende a gerar espaços semânticos mais dispersos para determinadas unidades lexicais, o que exige mais cuidado ao interpretar a MSL, uma vez que algumas das variações capturadas podem ser reflexo direto dessa heterogeneidade discursiva e não necessariamente de uma mudança lexical consolidada na língua.

Para além das limitações de arquitetura já discutidas no presente trabalho, os resultados obtidos por meio dos modelos contextuais, em especial, os obtidos com o conjunto de lexias β , apresentam informações que dialogam com a

Lexia	k em t_1		k em t_2	
	XLM-Roberta	BERTimbau	XLM-Roberta	BERTimbau
academia	2	2	2	2
homem	2	2	2	2
indústria	2	3	2	6
memória	2	2	2	3
mestre	2	8	2	2
negro	2	2	2	3
pecado	2	2	2	4
pessoa	2	2	2	4
rede	2	2	2	2
sala	2	5	2	4
virgem	2	2	2	10
lagarta	—	—	2	2
broca	—	—	2	3
biscate	—	—	2	2
carambola	2	2	2	4
mortal	2	2	2	3
perereca	—	—	2	2
peteca	—	—	8	10
birosca	—	—	3	3
estilete	—	—	2	9
ruge	—	—	2	2

Tabela 16: Valores de k identificados pelo método de silhueta e adotados para os conjuntos de lexias α e β em t_1 e t_2 . As células que não possuem valores, correspondem a unidades lexicais não encontradas nos dados da respectiva janela temporal (t_1 e t_2).

história e a lexicografia do português brasileiro. As trajetórias de unidades lexicais como “rede” e “memória” demonstram neologia semântica (a extensão do sentido de uma lexia motivada por inovações tecnológicas) (Alves, 1990). Além disso, as transformações observadas em “mestre” e “academia” refletem a influência das mudanças culturais e institucionais na permeabilidade do léxico (Biderman, 2001). A unidade lexical “mestre”, associado ao trabalho artesanal ou à figura do sábio, é dividido no *corpus* contemporâneo para designar majoritariamente um “grau acadêmico”, evidenciando a institucionalização da pós-graduação. Também, “academia”, antes restrito a centros filosóficos ou científicos, sofreu uma ampliação e posterior especialização no uso cotidiano para se referir a espaços de prática de atividade física.

Quanto à pejoração identificada na lexia “homem” na Janela B, cujos agrupamentos passaram a coocorrer com lexias que designam violência e crime, evidenciando uma mudança discursiva (de Moraes Pinto et al., 2024). Os resultados obtidos estão de acordo com a tipologia de mudança semântica (Ullmann, 1967), mas trazem implicações metodológicas para o estudo do português brasileiro. A mudança semântica não

decorre apenas de processos graduais de desgaste, mas é dinamicamente impulsionada por influências tecnológicas e sociais.

7. Conclusão e Trabalhos Futuros

O presente trabalho apresentou um método para a identificação e caracterização da Mudança Semântica Lexical (MSL) em língua portuguesa, utilizando *embeddings* de modelos estáticos e contextuais. Ao explorar múltiplas abordagens, com o treinamento de modelos estáticos e o ajuste fino de modelos contextuais em *corpora* diacrônicos (*corpus* ALiB-Twitter e *corpus* Tycholina), foram apresentadas características e limitações de cada método.

O principal avanço metodológico desta pesquisa para a língua portuguesa é ir além das análises que avaliam exclusivamente o sentido global e fixo das palavras. Ao organizar textos que cobrem séculos por meio do uso do *corpus* Tycholina, observamos que os modelos estáticos se mostraram eficazes em capturar mudanças semânticas globais, em contraste com os modelos contextuais, com o agrupamento *K-means*, que ofereceu uma identificação de *shifts* semânticos

mais sutis. Com isso, o estudo avança ao unir as representações computacionais dinâmicas à validação feita por linguistas, apresentando um método para rastrear a MSL.

Apesar da metodologia proposta mostrar-se capaz de sinalizar mudanças semânticas no léxico do português, ressalta-se a necessidade de uma melhoria na forma de identificação e interpretação dessas mudanças, em especial nos agrupamentos gerados por meio dos *embeddings* contextualizados, em relação à necessidade de suporte de avaliação qualitativa humana na leitura de amostras de sentenças do *corpus* para garantir a validade linguística dos agrupamentos.

Como proposta de avanços desta metodologia, futuras investigações devem se concentrar na otimização do agrupamento não supervisionado, visando a uma investigação sistemática dos hiperparâmetros do algoritmo. Propõe-se também a exploração de outras abordagens de agrupamento, a fim de comparar o desempenho e determinar qual seria o método mais robusto para a partição de *embeddings* lexicais contextuais. Além disso, a expansão do conjunto de lexias a ser investigado e o uso de *corpora* diacrônicos de outros domínios ou gêneros textuais, assim como a exploração da integração de representações fonéticas nos modelos contextuais para a investigação da variação fonológica (ou dialetal), da estabilidade e evolução do significado ao longo do tempo, são alguns dos potenciais trabalhos futuros a serem investigados.

Agradecimentos

Os autores agradecem o apoio do Projeto ALiB (Atlas Linguístico do Brasil), INCT-TILDIAR (#408490/2024-1), FAPESB (BOL0129/2025) e INCITE PIE0002/2022.

Referências

- Aitchison, Jean. 2002. *Language change: Progress or decay?* Cambridge University Press. [doi 10.1075/lplp.25.3.14roj](https://doi.org/10.1075/lplp.25.3.14roj)
- Alves, Ieda Maria. 1990. *Neologismo: criação lexical*. Atica
- Barranco, Roberto Camacho, Raimundo F. Dos Santos, M. Shahriar Hossain & Monika Akbar. 2018. Tracking the evolution of words with time-reflective text representations. Em *IEEE International Conference on Big Data*, 2088–2097. [doi 10.1109/BigData.2018.8621902](https://doi.org/10.1109/BigData.2018.8621902)

- Beck, Christin. 2020. DiaSense at SemEval-2020 task 1: Modeling sense change via pre-trained BERT embeddings. Em *14th Workshop on Semantic Evaluation*, 50–58. [doi 10.18653/v1/2020.semeval-1.4](https://doi.org/10.18653/v1/2020.semeval-1.4)
- Biderman, Maria Tereza Camargo. 2001. Terminologia e lexicografia. *Tradterm* 7. 153–181. [doi 10.11606/issn.2317-9511.tradterm.2001.49147](https://doi.org/10.11606/issn.2317-9511.tradterm.2001.49147)
- Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017a. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)* 5. 135–146. [doi 10.1162/tacl_a_00051](https://doi.org/10.1162/tacl_a_00051)
- Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017b. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)* 5. 135–146. [doi 10.1162/tacl_a_00051](https://doi.org/10.1162/tacl_a_00051)
- Cardoso, Suzana & Jacyra Mota. 2014. *Atlas linguístico do brasil*. Addison-Wesley
- Cardoso, Suzana, Jacyra Mota, Vanderci Aguilera, Maria do Socorro de Aragão, Aparecida Isquendo, Abdelhak Razky, Felício Margotti & Cléo Altenhofen. 2014a. *Atlas linguístico do brasil*, vol. 1. Londrina: Eduel
- Cardoso, Suzana, Jacyra Mota, Vanderci Aguilera, Maria do Socorro de Aragão, Aparecida Isquendo, Abdelhak Razky, Felício Margotti & Cléo Altenhofen. 2014b. *Atlas linguístico do brasil*, vol. 2. Londrina: Eduel
- Coelho, Izete Lehmkuhl, Edair Gorski, Christiane Maria Nunes de Souza & Guilherme Henrique May. 2015. *Para conhecer sociolinguística*. Editora Contexto. [doi 10.18653/v1/2020.acl-main.747](https://doi.org/10.18653/v1/2020.acl-main.747)
- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. Em *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 8440–8451. [doi 10.18653/v1/2020.acl-main.747](https://doi.org/10.18653/v1/2020.acl-main.747)
- Crespo, Maria Clara Ramos Morales, Maria Lina de Souza Jeannine Rocha, Mariana Lourenço Sturzeneker, Felipe Ribas Serras, Guilherme Lamartine de Mello, Aline Silva Costa, Mayara Feliciano Palma, Renata Moraes Mesquita, Raquel de Paula Guets, Mariana Marques da Silva, Marcelo Finger, Maria Clara Paixão de Sousa, Cristiane Namiuti

- & Vanessa Martins do Monte. 2023. Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information. ArXiv [cs.CL]. doi 10.48550/arXiv.2303.16098
- Crowley, Terry & Claire Bown. 2010. *An introduction to historical linguistics* chap. 11. Oxford University Press
- De Saussure, Ferdinand. 2006. *Curso de lingüística geral*. Editora Cultrix 27th edn.
- Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. Em *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 457–470. doi 10.18653/v1/P19-1044
- Fonseca Lage, Lucas & Evandro Landulfo Teixeira Paradelo Cunha. 2022. Mudança semântica e word embeddings: estudos de caso na diacronia do português. *Revista de Estudos da Linguagem* 30(4). 2043–2086. doi 10.17851/2237-2083.30.4.2043-2086
- Galves, Charlotte. 2018. The tycho brahe corpus of historical Portuguese: Methodology and results. *Linguistic Variation* 18(1). 49–73. doi 10.1075/lv.00004.gal
- Hamilton, William, Jure Leskovec & Dan Jurafsky. 2016a. Diachronic word embeddings reveal statistical laws of semantic change. Em *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1489–1501. doi 10.18653/v1/P16-1141
- Hamilton, William L, Jure Leskovec & Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. Em *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1489–1501. doi 10.18653/v1/P16-1141
- Hilpert, Martin & Stefan Th. Gries. 2016. Quantitative approaches to diachronic corpus linguistics. Em *The Cambridge handbook of English historical linguistics*, chap. 2, 36–53. Cambridge University Press. doi 10.1017/CB09781139600231.003
- Joulin, Armand, Edouard Grave, Piotr Bojanowski & Tomas Mikolov. 2017a. Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* 2. 427–431. [↗](#)
- Joulin, Armand, Edouard Grave, Piotr Bojanowski & Tomas Mikolov. 2017b. Bag of tricks for efficient text classification. Em *15th European Chapter of the Association for Computational Linguistics (EACL)*, 427–431. [↗](#)
- Kutuzov, Andrey & Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. Em *14th Workshop on Semantic Evaluation*, 126–134. doi 10.18653/v1/2020.emeval-1.14
- Levy, Omer & Yoav Goldberg. 2014. Dependency-based word embeddings. Em *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 302–308. doi 10.3115/v1/P14-2050
- Martinc, Matej, Petra Kralj Novak & Senja Polak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. Em *20th Language Resources and Evaluation Conference (LREC)*, 4811–4819. [↗](#)
- Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv [cs.CL]. doi 10.48550/arXiv.1301.3781
- Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. Em *1st International Conference on Learning Representations (ICLR)*, [↗](#)
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. Em *Advances in Neural Information Processing Systems*, [↗](#)
- de Moraes Pinto, Deise Cristina, Fábio André Cardoso Coelho, Mônica Paula de Lima Cabral & Roza Maria Palomanes Ribeiro. 2024. *Introdução à semântica*. Fundação CECIERJ
- Řehůřek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. Em *LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. [↗](#)
- Ruas, Gabrielle. 2023. O item lexical “lacrar” nas redes sociais: Uma análise diacrônica. *Calígrama: Revista de Estudos Românicos* 28(1). 26–63. doi 10.17851/2238-3824.28.1.26-63
- Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. Em *14th Workshop on Semantic Evaluation*, 1–23. doi 10.18653/v1/2020.emeval-1.1

- Silva, Rosa Virgínia Mattos. 2018. Teorias da mudança lingüística e a sua relação com a(s) história(s) da(s) língua(s). *Linguística: Revista de Estudos Linguísticos da Universidade do Porto* 3. 39–53. [↗](#)
- Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. Em *9th Brazilian Conference on Intelligent Systems (BRACIS)*, [doi](#) [10.1007/978-3-030-61377-8_28](#)
- Sturzeneker, Mariana, Maria Clara Crespo, Maria Lina Rocha, Marcelo Finger, Maria Clara Paixão de Sousa, Vanessa Martins do Monte & Cristiane Namiuti. 2022. Carolina's methodology: building a large corpus with provenance and typology information. Em *2nd Workshop on Digital Humanities and Natural Language Processing (DHandNLP)*, 53–58. [↗](#)
- de Sá, Jader Martins Camboim, Marcos Da Silveira & Cédric Pruski. 2026. Survey in characterization of semantic change. *Natural Language Processing* 32(3). [doi](#) [10.1017/nlp.2026.10023](#)
- Ullmann, Stephen. 1967. *Semântica: uma introdução à ciência de significado* Manuais universitários. Fundação Calouste Gulbenkian. [↗](#)