

EACL 2026

**The 6th International Workshop on Computational
Approaches to Language Change (LChange'26)**

Proceedings of the Workshop

March 28, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-362-3

Preface by the General Chair

Welcome to the 6th International Workshop on Computational Approaches to Language Change (LChange'26) co-located with EACL 2026. After a year-long break, we are now back with a highly anticipated edition. LChange is held on March 28th, 2026, as a half-day workshop where the primary mode of presentation is poster presentation. The introduction and lightning talk sessions are offered in hybrid mode, while the poster session is on-site in Morocco. Because we were only allotted a half-day workshop, for the first time in LChange history, there will be no keynotes. Our keynotes have been very appreciated in the past and we intend to have them back in the next edition. New for this year is that we have dropped the word *Historical* from the title, to highlight that we are open to the study of all language change, modern and historical alike.

Characterizing the time-varying nature of language will have broad implications and applications in multiple fields, including linguistics, artificial intelligence, digital humanities, computational cognitive and social sciences. In this workshop, we bring together the world's pioneers and experts in computational approaches to language change with a focus on digital text corpora. In doing so, this workshop carries out the triple goals of disseminating state-of-the-art research on diachronic modeling of language change, fostering cross-disciplinary collaborations, and exploring the fundamental theoretical and methodological challenges in this growing niche of computational linguistic research.

In response to the call, we received 24 submissions of which two were desk rejected because they did not fit the workshop theme. Each of the remaining submissions was carefully evaluated by at least two members of the Program Committee, whom we believed to be most appropriate for each paper. Most papers received 3 or more reviews. Based on the reviewers' feedback, we accepted 14 full and short papers as poster presentations. This year, we also received one submission that we deemed to be (fully or majorly) LLM-generated, which was rejected by the organizers after careful reviewing. The 14 accepted papers cover a wide range of topics, phenomena and methods in the field of language change. The methods employed span deep neural networks with monolingual and multilingual language models, and statistical analyses based on usage similarities and dependency co-occurrence patterns. We also highlight the diversity of cultural and historical contexts for language change studied in these papers, featuring studies on Japanese economic language during the "Lost Decade" (1990–2000), tonogenesis in Tibetan, and slang word detection. Additionally, some papers focus on contemporary sociolinguistic phenomena, such as neology in social media or semantic broadening in specific contexts, while others introduce tools and resources for modeling semantic change on a broad scale. This diversity underscores the richness of current research on language change.

We hope that you will find the workshop papers insightful and inspiring. We would like to thank the authors of all papers for their interesting contributions, and the members of the Program Committee for their insightful reviews. Our special thanks go to the emergency reviewers who stepped in to provide their expertise. We also express our gratitude to the EACL 2026 workshop chairs for their kind assistance during the organization process. Finally, our thanks go to our sponsors, the research program "Change is Key!" (Riksbankens Jubileumsfond, contract M21-0021).

As in the past editions, we are looking forward to a productive workshop and hope that it will foster new collaborations and contribute to building an ever-growing community.

Nina Tahmasebi, chair, University of Gothenburg (Sweden)

Pierluigi Cassotti, University of Gothenburg (Sweden)

Andrey Kutuzov, University of Oslo (Norway)

Netta Huebscher, University of Gothenburg (Sweden)

Elena Spaziani, Sapienza, University of Rome (Italy)

Naomi Baes, Melbourne School of Psychological Sciences, University of Melbourne (Australia)

David Alfter, University of Gothenburg (Sweden)

Syrielle Montariol, CNRS - Sorbonne Université (France)

LChange'26 Workshop Chairs

Organizing Committee

General Chair

Nina Tahmasebi, University of Gothenburg, Sweden

Workshop Chairs

Pierluigi Cassotti, University of Gothenburg, Sweden

Andrey Kutuzov, University of Oslo, Norway

Netta Huebscher, University of Gothenburg, Sweden

Elena Spaziani, Sapienza University of Rome, Italy

Naomi Baes, University of Melbourne, Australia

David Alfter, University of Gothenburg, Sweden

Syrielle Montariol, Sorbonne University, France

Program Committee

Program Chairs

Naomi Baes, University of Melbourne
Pierluigi Cassotti, Göteborg University
Netta Huebscher, Göteborg University
Andrey Kutuzov, University of Oslo
Syrielle Montariol, Sorbonne University
Elena Spaziani, University of Rome
Nina Tahmasebi, Göteborg University

Program Committee

Taichi Aida, Özge Alacam

Naomi Baes, Pierpaolo Basile, Frederic Blum

Kalvin Chang, Jing Chen

Stefano De Pascale, Aleksei Dorkin

Mariia Fedorova, Pascale Feldkamp, Karlien Franco

Mario Giulianelli, Roksana Goworek, Mauricio Gruppi

Johann-Mattis List

Barbara McGillivray, Timothee Mickus, Syrielle Montariol, Tony Montes, David R. Mortensen, Pablo Mosteiro

Bill Noble, Krzysztof Nowak

Lidia Pivovarova

Zhengfei Ren

Yves Scherrer, Weiwei Sun

Nina Tahmasebi, Ludovic Tanguy, Niko Tapio Partanen, Stephen Eugene Taylor

Table of Contents

<i>The SlangTrack Dataset: Supporting the Detection of Words Used in Slang Senses</i> Afnan Mohammed Aloraini, Riza Batista-Navarro, Goran Nenadic and Viktor Schlegel	1
<i>Statistical Semantic Change Detection via Usage Similarities</i> Taichi Aida, Daichi Mochihashi, Hiroya Takamura, Toshinobu Ogiso and Mamoru Komachi . .	20
<i>The Tonogenesis Continuum in Tibetan: A Computational Investigation</i> Siyu Liang and Zhaxi Zerong	27
<i>Cross-lingual Lexical Semantic Change in Romance Languages</i> Ana Sabina Uban, Liviu P Dinu, Anca Daniela Dinu and Simona Georgescu	35
<i>Threshold-Calibrated Word Sense Disambiguation: Semantic Broadening Without Sense Redistribution in Schizophrenia</i> Naomi Baes and Nick Haslam	50
<i>Using Correspondence Patterns to Identify Irregular Words in Cognate Sets Through Leave-One-Out Validation</i> Frederic Blum and Johann-Mattis List	75
<i>DHPLT: large-scale multilingual diachronic corpora and word representations for semantic change modelling</i> Mariia Fedorova, Andrey Kutuzov and Khonzoda Umarova	87
<i>Transparent Semantic Change Detection with Dependency-Based Profiles</i> Bach Phan Tat, Kris Heylen, Dirk Geeraerts, Stefano De Pascale and Dirk Speelman	97
<i>Semantic Change Characterization with LLMs using Rhetorics</i> Jáder Martins Camboim de Sá, Jooyoung Lee, Marcos Da Silveira and Cedric Pruski	110
<i>Using BERT to Explore Lexical Semantic Change of Prepositions</i> Liudmila Radchankava	124
<i>A Computational Analysis of the Emergence of Therapy-speak in Social Media</i> Alina Iacob and Ana Sabina Uban	131
<i>Rethinking Metrics for Lexical Semantic Change Detection</i> Roksana Goworek and Haim Dubossarsky	147
<i>From sunblock to softblock: Analyzing the correlates of neology in published writing and on social media</i> Maria Ryskina, Matthew R. Gormley, Kyle Mahowald, David R. Mortensen, Taylor Berg-Kirkpatrick and Vivek Kulkarni	162

Program

Saturday, March 28, 2026

09:15 - 09:45 *Introduction*

09:45 - 10:30 *Poster pitch*

10:30 - 11:00 *Coffee Break*

10:30 - 12:30 *Poster Session*

Threshold-Calibrated Word Sense Disambiguation: Semantic Broadening Without Sense Redistribution in Schizophrenia

Naomi Baes and Nick Haslam

From sunblock to softblock: Analyzing the correlates of neology in published writing and on social media

Maria Ryskina, Matthew R. Gormley, Kyle Mahowald, David R. Mortensen, Taylor Berg-Kirkpatrick and Vivek Kulkarni

DHPLT: large-scale multilingual diachronic corpora and word representations for semantic change modelling

Mariia Fedorova, Andrey Kutuzov and Khonzoda Umarova

Using BERT to Explore Lexical Semantic Change of Prepositions

Liudmila Radchankava

Using Correspondence Patterns to Identify Irregular Words in Cognate Sets Through Leave-One-Out Validation

Frederic Blum and Johann-Mattis List

Transparent Semantic Change Detection with Dependency-Based Profiles

Bach Phan Tat, Kris Heylen, Dirk Geeraerts, Stefano De Pascale and Dirk Speelman

Rethinking Metrics for Lexical Semantic Change Detection

Roksana Goworek and Haim Dubossarsky

The SlangTrack Dataset: Supporting the Detection of Words Used in Slang Senses

Afnan Mohammed Aloraini, Riza Batista-Navarro, Goran Nenadic and Viktor Schlegel

Saturday, March 28, 2026 (continued)

Tracking Lexical Semantic Change with a Three-Level Network: Evidence from Japanese Economic News during Japan's Lost Decade"(1990-2000)

Jiabao Wang

The Tonogenesis Continuum in Tibetan: A Computational Investigation

Siyu Liang and Zhaxi Zerong

Statistical Semantic Change Detection via Usage Similarities

Taichi Aida, Daichi Mochihashi, Hiroya Takamura, Toshinobu Ogiso and Mamoru Komachi

Semantic Change Characterization with LLMs using Rhetorics

Jáder Martins Camboim de Sá, Jooyoung Lee, Marcos Da Silveira and Cedric Pruski

Cross-lingual Lexical Semantic Change in Romance Languages

Ana Sabina Uban, Liviu P Dinu, Anca Daniela Dinu and Simona Georgescu

A Computational Analysis of the Emergence of Therapy-speak in Social Media

Alina Iacob and Ana Sabina Uban

The SlangTrack Dataset: Supporting the Detection of Words Used in Slang Senses

Afnan Aloraini^{1,2}, Goran Nenadic¹, Viktor Schlegel¹ and Riza Batista-Navarro¹

¹Department of Computer Science, The University of Manchester, UK

²Department of Computer Science, Qassim University, Saudi Arabia

{afnan.aloraini, g.nenadic, viktor.schlegel, riza.batista}@manchester.ac.uk

A.ALOURANI@qu.edu.sa

Abstract

Slang is widespread in informal communication, yet its fluidity poses challenges for natural language processing (NLP), especially when words alternate between slang and non-slang senses. While prior work has examined slang through dictionaries, sentiment analysis, and lexicon building, little attention has been given to detecting slang usage in context. We address this gap by re-framing slang detection as distinguishing slang from non-slang senses of the same lexical item. To support this task, we introduce *SlangTrack* (ST), a diachronically structured dataset of dual-meaning words annotated at the sentence level with high inter-annotator agreement. We benchmark (1) deep learning models with static and contextual embeddings, (2) transformer-based models, and (3) large language models evaluated in zero-shot, few-shot, and fine-tuned settings. Fine-tuned transformers, especially BERT-large enriched with sentiment and emotion features, achieve the strongest performance, reaching an F1-score of 72% for slang and 92% for non-slang usage. Our findings highlight both the difficulty of contextual slang detection and the value of affective cues for improving model robustness.

1 Introduction

Disclaimer: This work includes offensive slang examples, which do not reflect the researchers' views.

Slang is an informal linguistic phenomenon comprising words and phrases used within specific groups (Dumas and Lighter, 1978; Adams, 2012; Green, 2015; Sun et al., 2021). Although linguists and computer scientists have developed many methods for semantic analysis, slang remains a persistent challenge for natural language processing (NLP) systems (Eisenstein, 2013; Blodgett et al., 2016; Wuraola et al., 2024). It often encodes emotions, attitudes, and social affiliation elements that are crucial for interpreting meaning in context (Sandow et al., 2024; Haber and Poesio, 2024).

Identifying slang is important for tracking semantic change and improving NLP tasks (Adams, 2012; Sun et al., 2022; Keidar et al., 2022; Sun et al., 2024). Many terms evolve rapidly, for example, “cool” shifted from a literal descriptor to an evaluative slang term (Coleman, 2012; Dhuliawala et al., 2016). Accurate disambiguation of such meanings is essential for applications like conversational agents, machine translation, and sentiment analysis, where figurative senses are often misinterpreted.

Detecting slang is challenging because many terms have multiple meanings depending on context, often resulting in double entendres (Kiddon and Brun, 2011). For example, “He’s a player on Sundays” can refer to sports or manipulative romantic behaviour, highlighting the need for contextual disambiguation. Recent work in NLP has begun to recognise the importance of informal and slang language as meaningful phenomena for modern language technologies. For example, benchmarks and evaluation frameworks have been proposed to measure models’ capabilities in processing slang and other informal language forms using neural methods (Sun et al., 2024; Pei et al., 2019). Studies of large language models also show that slang remains challenging for foundational models and serves as a useful probe of their linguistic knowledge (Sun et al., 2024).

Despite these emerging efforts, slang detection has received relatively limited attention in the broader NLP community. Existing resources tend to focus either on dictionary–text mismatches (Pei et al., 2019) or on scripted conversational data such as movie subtitles (Sun et al., 2024), and do not address words that appear in both slang and literal senses. We refer to these as *dual-meaning words*.

In this work, slang detection is framed as a context-sensitive and pragmatic distinction between slang and non-slang usage of the same lexical item, rather than a simple lexical lookup. Be-

cause our analysis draws on both historical and contemporary corpora, it aligns with computational research on semantic change (Hamilton et al., 2016; Schlechtweg et al., 2020; Kutuzov et al., 2018; Periti et al., 2024). Many slang senses emerge as pragmatic extensions of earlier meanings, making dual-meaning words a useful lens for studying sense competition and semantic–pragmatic drift over time (de Sá et al., 2024; Tahmasebi et al., 2021). This framing is inspired by word sense disambiguation (WSD) research, without committing to full sense-level modelling, and allows register-level (slang vs. non-slang) annotations to be aligned across corpora in a controlled manner.

This paper introduces a binary classification system designed to determine whether a target word occurrence is used in a slang or non-slang (standard) sense within its given textual instance. Our primary research question is: *Can algorithms reliably distinguish slang usage from non-slang usage at the instance level?* For example, the word “salty” can be used literally (“the soup is salty”) or figuratively to express resentment (“she was salty after losing”). We frame slang detection as an instance-level (target-word-centered) classification task, where each instance is annotated and evaluated individually. Each instance corresponds to the same textual input provided to both annotators and models: a single tweet (which may contain multiple sentences) in Twitter, or a paragraph-length context in CCOHA. This framing provides a controlled and comparable unit of analysis across corpora, while allowing naturally occurring pragmatic and contextual cues to be captured within the textual span itself. Our contributions include the following:

- SlangTrack, a new corpus annotated with slang and non-slang labels at the instance level, supporting binary classification of target-word usage.
- Benchmarking a range of slang detection models, including: (1) basic neural classification models, (2) fine-tuned transformer-based language models (LMs), and (3) large language models (LLMs) evaluated under fine-tuned, zero-shot, and few-shot settings.
- An analysis of the role of sentiment and emotion cues in distinguishing slang from non-slang usage.
- An error analysis of misclassified instances produced by the best-performing model.

2 Related work

2.1 Construction of Slang Dictionaries and Sentiment Analysis

The development of resources for processing informal language, particularly slang, has been central in computational linguistics. Early efforts focused on constructing structured representations of slang, such as SlangNet (Dhuliawala et al., 2016), which integrates Urban Dictionary¹ entries into a WordNet-style framework (Miller, 1995). SlangSD (Wu et al., 2018) extends this line of work by providing a continually updated sentiment lexicon for slang terms derived from large-scale web and social media data. SLANGZY (Gupta et al., 2019) further applies machine learning to dynamically interpret slang and support downstream applications such as chatbots and social media analytics. Although these resources expand lexical and sentiment coverage, they do not annotate contextual usage or distinguish between non-slang and slang senses of the same word, which is central to our task.

2.2 Slang Word Creation and Interpretation

Research on slang detection and interpretation has evolved alongside the growth of digital communication. Early studies constructed slang lexicons (Pal and Saha, 2015), while later work employed neural architectures such as BiLSTMs, CRFs, and MLPs for classification (Pei et al., 2019). Beyond detection, sequence-to-sequence models have been used to interpret non-standard English (Ni and Wang, 2017), and BiLSTM- and GRU-based systems have shown strong performance in identifying domain-specific slang (Lynn et al., 2019). More recently, the OpenSub-Slang dataset (Sun et al., 2024) introduced demographic, contextual, and historical metadata for 7,488 slang-related sentences, enabling more nuanced modelling. Additional work combining semantic and contextual cues (Sun et al., 2022) has further advanced slang interpretation.

Recent developments in WSD are also relevant to our task. Models such as GlossBERT (Huang et al., 2019), EWISE (Kumar et al., 2019), and EWISER (Bevilacqua and Navigli, 2020) demonstrate the effectiveness of supervised context–gloss methods for resolving semantic ambiguity. However, these approaches do not target slang senses specifically or model the contrast between slang and literal usage, which remains understudied.

¹<https://www.urbandictionary.com>

2.3 Research Gaps

Although prior work has advanced slang detection, it has generally treated slang as a broad category rather than focusing on dual-meaning words. For example, Pei et al. (2019) contrasts dictionary-derived slang with formal news text, and Sun et al. (2024) captures conversational slang using movie subtitles. Other studies draw on Urban Dictionary or focus on specific domains such as misogynistic slang (Lynn et al., 2019). More recently, the OpenSubtitles-Slang dataset introduced cross-lingual mappings for slang terms, providing paraphrases or equivalents in multiple languages (Sun et al., 2024). However, in these resources, slang and non-slang items originate from different corpora, so the same lexical item is not annotated with both slang and non-slang senses. As a result, the task becomes identifying slang words as a category rather than determining whether a familiar word is used in a slang or a non-slang sense in context.

Our study addresses this gap by re-framing slang detection as an instance-level contextual disambiguation problem. Instead of asking whether a text contains slang expressions, we ask whether a specific occurrence of a polysemous word is used with a slang (socially marked/figurative) or non-slang (literal/standard) meaning in context. This perspective is inspired by word sense disambiguation, but focuses on register-level distinctions, extending prior slang research to an underexplored setting in which the same lexical item alternates between slang and non-slang usage.

3 Problem Formulation

We frame slang detection as an instance-level classification task inspired by word sense disambiguation, applied to polysemous dual-meaning words, each with at least one slang and one non-slang sense. Let w denote such a target word and s a sentence containing w . Each instance is annotated with a binary label indicating whether its sense in context is slang or non-slang, and the task is to assign the correct label $y \in \{\text{slang}, \text{non-slang}\}$ to each occurrence of w .

The word *salty* illustrates the range of senses we consider: a literal taste meaning (“the soup is salty”), slang for resentment (“he was salty after losing”), slang for old or worn (“a salty jacket”), slang for toughness (“a salty veteran”), slang for vulgarity (“a salty bar”), and non-slang proper-noun uses (“Salty is the name of their dog”). These

senses demonstrate how a single lexical item spans multiple non-slang and slang sub-senses; in SlangTrack, this variability is intentionally collapsed into a binary decision, making contextual disambiguation of socially marked (slang) versus non-slang usage central to the task. This formulation prioritises register-level discrimination over fine-grained sense distinctions.

4 Dataset

Existing resources for slang research face limitations for binary slang classification and polysemy-sensitive analysis. Dictionary-based resources such as Urban Dictionary (Ni and Wang, 2017), the Online Slang Dictionary (OSD),² and Green’s Dictionary of Slang (GDoS) (Adams, 2012) offer broad lexical coverage but lack systematically annotated non-slang counterparts and do not capture contextual usage. Reddit glossaries (Lucy and Bamman, 2021) provide community-specific slang terms, but remain at the glossary level rather than sentence level.

Two datasets support binary slang classification: Pei et al. (2019) and OpenSub-Slang (Sun et al., 2024). Pei et al. combine dictionary-derived slang with negative samples drawn heuristically from news text, raising the risk of domain artefacts. OpenSub-Slang offers richer contextual information, including paraphrases and demographic metadata, but its reliance on scripted dialogue introduces domain biases and limits generalisation to real-world slang. Neither resource addresses dual-meaning words that alternate between slang and non-slang meanings. A comparison of slang-related datasets is shown in Table 1.

To address these gaps, we introduce the SlangTrack (ST) dataset.³ The ST dataset is built around dual-meaning target words, each possessing at least one slang and one non-slang sense. Unlike dictionary-based or scripted sources, ST draws on naturally occurring language in both the Cleaned Corpus of Historical American English (CCOHA) (Alatrash et al., 2020) and contemporary Twitter.⁴ This combination provides coverage of modern slang usage while enabling contextual

²<https://www.onlineslangdictionary.com>

³Publicly available at: <https://github.com/SlangTrack/SlangTrack-ST/blob/main/README.md>

⁴<https://twitter.com>. Our dataset includes social media text, which may contain offensive material. All excerpts are de-identified in accordance with Twitter’s Academic Research TOS.

Dataset	Source / Domain	Temp. Cov.	Slang Cov.	Non-slang Cov.	Polysemy (dual-meaning)	Annotation	Fully Annot.	Bench.	Publicly Available
<i>Urban Dictionary</i> (Ni and Wang, 2017)	Crowdsourced slang dictionary (defs. + examples)	Contemp.	✓	✗	✗	Word / entry-level	✗	✗	✓
<i>Online Slang Dictionary (OSD)</i> (Sun et al., 2022)	Curated slang dictionary (defs. + examples)	Contemp.	✓	✗	✗	Word / entry-level	✗	✗	✗
<i>Green’s Dictionary of Slang (GDoS)</i> (Adams, 2012)	Historical lexicographic dictionary (defs. + citations)	Hist.	✓	✗	✗	Word / entry-level	✗	✗	✗
<i>Reddit Glossaries</i> (Lucy and Bamman, 2021)	Community slang glossaries (subreddit lists)	Contemp.	✓	✗	✗	Word / glossary-level	✗	✗	✓
<i>OpenSubtitles-Slang</i> (Sun et al., 2024)	Scripted dialogue (movie subtitles)	Contemp.	✓	✓	✗	Sentence-level (+ slang tokens, subset annotated)	✗	✓	✓
<i>SlangTrack (ST)</i>	Naturally occurring text in two diverse corpora	Both	✓	✓	✓	Instance-level (target-word-centered; slang vs. non-slang)	✓	✓	✓

Table 1: Comparison of slang-related datasets. “Temp. Cov.” = temporal coverage (“Hist.” pre-2000; “Contemp.” post-2000; “Both” spans both). “Slang / Non-slang Cov.” = slang vs. non-slang usage. “Polysemy” = dual meanings. “Annotation” = granularity. “Fully Annot.” = fully gold-labelled. “Bench.” = benchmark.

disambiguation across time and registers.

4.1 Data Collection (Target Words and Examples)

We selected target words that appeared in both the SlangSD wordlist⁵ and CCOHA, prioritising items with multiple attested senses. Each chosen word has at least one slang and one non-slang sense. To compile sense inventories, we consulted Green’s Dictionary of Slang, Urban Dictionary, and the Online Slang Dictionary, and cross-referenced meanings with the Oxford English Dictionary (OED)⁶ to confirm non-slang usage. Full sense inventories are provided in Appendix 12.

We ensured that each word appeared in both CCOHA and Twitter, allowing the dataset to reflect historical and contemporary senses. Applying these criteria yielded ten target words. Although ST contains only ten target words, this limited lexical scope is intentional and methodologically motivated. Each selected word exhibits between two and eight attested senses, requiring careful cross-referencing across slang dictionaries, OED entries, and corpus attestations. Because contextual disambiguation for dual-meaning words is annotation-intensive, especially for long and paraphrastic CCOHA sentences, focusing on a smaller set ensured high annotation quality and interpretability of sense contrasts (see Appendix 16). This design follows established practice in semantic change research, where controlled, high-quality

pilot datasets with narrow lexical coverage (e.g., SemEval 2020 Task 1) serve as foundational benchmarks before broader scaling. Accordingly, ST itself should be interpreted as a controlled pilot dataset, prioritising sense precision and annotation quality over breadth. This initial release establishes a reliable foundation that we will expand in future work with additional words, platforms, and more recent data. Each instance is labelled with its source (CCOHA or Twitter) and timestamp (CCOHA publication year or tweet date). An instance corresponds to the naturally occurring textual unit in the source corpus: a single tweet in Twitter or a paragraph-length context in CCOHA. While instances are evaluated independently, their length varies across sources, allowing some pragmatic and contextual cues to be captured without explicitly modelling broader discourse history. Although the task centres on slang detection rather than temporal modelling, this metadata enables users to distinguish older formal contexts from modern informal ones. While final labels are binary (slang vs. non-slang), annotators relied on the predefined sense inventory to determine whether each instance corresponded to an attested slang or non-slang sense. This ensures that the slang label reflects a genuine register distinction grounded in attested meanings, rather than surface stylistic informality.⁷

⁷We also release a fine-grained sense-labelled version, SlangTrack-WSD (ST-WSD), derived from the same annotation process. While ST-WSD preserves fine-grained sense distinctions, the present study focuses on binary register-level disambiguation; a direct comparison between binary and

⁵<https://rdocumentation.org/packages/lexicon>

⁶<https://www.oed.com>

Taken together, these design decisions give ST synchronic labels over a diachronic dataset, where slang and non-slang senses show distinct patterns across CCOHA and Twitter. Even without explicit temporal modelling, this structure enables analysis of how socially marked senses vary across time and registers.

CCOHA contributed historical usage spanning 1980–2010, including non-slang and occasional slang contexts. For contemporary usage, we collected up to 1,000 tweets per target word from 2010–2020 through the Twitter API. Combining both corpora allows ST to capture emerging slang from social media alongside formal usage from CCOHA, balancing sources and enabling direct comparison of dual-meaning words across registers. We deliberately restrict our sources to CCOHA and Twitter to create a clear contrast between historical formal writing and contemporary informal usage.

4.2 Annotation Guidelines and Details

Annotators were provided with the target words, example sentences, and a predefined sense inventory. Using this information, they labelled each instance as *slang* or *non-slang* based on whether the usage matched an attested slang sense or a non-slang sense.⁸ All instances were mapped to a binary slang versus non-slang classification setting.

The annotation team consisted of three English-proficient annotators, one with a linguistics background who served as the primary annotator. Prior to full-scale annotation, an initial pilot phase was conducted to refine the annotation guidelines. Following this, two annotators independently labelled all instances in the dataset (12,712 instances) using the finalised guidelines.

Inter-annotator agreement was computed on the independent (pre-adjudication) labels using Cohen’s Kappa and reached 0.887, indicating high overall reliability. Disagreements were subsequently resolved by the primary annotator through adjudication. The resulting class distribution is

sense-level labels is left to future work. ST-WSD is provided as a companion resource: <https://github.com/SlangTrack/SlangTrack-Word-Sense-Disambiguation>.

⁸Proper-noun usages (e.g., brand names, organization names, song titles) are grouped under the non-slang category. While these uses are not always semantic “senses” in a strict lexicographic sense, they are not socially marked or figurative in the way slang usages are. For the purposes of this task, which focuses on distinguishing slang from non-slang register rather than enumerating fine-grained sense inventories, treating proper-noun usages as non-slang provides a consistent and operational distinction.

skewed, with approximately 80% non-slang and 20% slang instances. This imbalance reflects naturally occurring usage patterns rather than annotation or sampling decisions. To account for class imbalance, we report macro-averaged evaluation metrics and conduct per-word analyses rather than relying on accuracy alone.

4.3 Data Statistics

The dataset contains 12,712 labelled instances: 10,105 non-slang and 2,607 slang. It includes 48,508 unique word types and 310,170 tokens. The average post length is 34.6 words and contains 3.74 sentences. Using stratified sampling, the dataset was divided into 70% training, 15% validation, and 15% test splits.

Keyword	Non-slang	Slang	Total
BMW	1,082	15	1,097
Brownie	706	258	964
Chronic	1,259	426	1,685
Climber	505	137	642
Cucumber	978	73	1,051
Eat	2,708	324	3,032
Germ	753	79	832
Mammy	904	166	1,070
Rodent	744	329	1,073
Salty	535	731	1,266
Total	10,105	2,607	12,712

Table 2: Distribution of slang and non-slang instances per target word.

Breaking this down by corpus, CCOHA contributes 3,660 non-slang and 756 slang examples, while Twitter contributes 6,514 non-slang and 1,782 slang examples. The resulting class distribution (approximately 80% non-slang and 20% slang) reflects naturally occurring usage patterns in the underlying corpora rather than annotation or sampling decisions.

To explore stylistic variation between slang and non-slang usage, we conducted a one-way ANOVA on automatically derived sentiment and emotion scores. Because these scores are produced by an external classifier, the analysis is exploratory and not used to draw causal conclusions or justify model performance. The results show statistically significant but small differences: slang instances tend to be slightly more negative and express anger and sadness, while non-slang instances are more neutral. Full descriptive statistics appear in Appendix 11.

4.4 Diachronic Variation in Slang and non-slang Senses

Because ST dataset includes both historical COHA data and contemporary Twitter data, it allows us to examine how the balance between slang and non-slang senses changes over time. In COHA, most target words appear predominantly in their literal senses (e.g., *Eat* <3% slang usage, *BMW* <6%, *Cucumber* <13%), while a few items such as *Rodent* and *Chronic* show moderate slang usage (around 25–30%). Twitter displays a different distribution. Some words become substantially more slang-dominant: *Salty*, for instance, shifts from being mostly literal in COHA to being used slang overwhelmingly in Twitter (>70%). Other items show the opposite pattern. *Mammy* appears frequently in slang senses in COHA, but these uses become rare in modern Twitter data. Several words, including *Eat*, *Climber*, and *Cucumber*, remain largely stable, with literal senses prevailing in both corpora. Although ST spans multiple historical periods, time is not modelled as an explicit predictive variable. Instead, it is treated as a latent property of the data: instances are timestamped, but classification relies solely on local textual context, allowing us to isolate instance-level slang disambiguation while still supporting descriptive temporal analysis distributed across periods.

A statistical comparison of slang proportions between COHA and Twitter (Appendix 13) confirms these trends. *Salty* and *Mammy* exhibit large, highly significant shifts ($p < .001$), while most other words show little or no significant change. Taken together, these results highlight that the dual-meaning words in ST follow diverse diachronic paths, illustrating patterns of sense innovation, decline, and stability over time.

5 Methodology

5.1 Pre-processing

We applied standard text pre-processing to reduce noise and ensure consistent input across models. This included removing URLs and usernames, excluding instances in which the target word appeared inside a URL, and eliminating duplicated content. Duplicate removal was performed at the exact string level, meaning that an instance was removed only if the entire textual content was repeated verbatim more than once in the dataset. Text normalisation steps such as lowercasing and punctuation removal were applied only in settings using

static word embeddings with CNN and BiLSTM models, where such preprocessing is standard practice. For transformer-based and large language models, text was minimally processed, preserving original casing and punctuation in line with their pre-training regimes.

5.2 Evaluation

Models were evaluated using precision, recall, F1 (macro and weighted), and accuracy. Macro scores emphasise performance on the minority slang class, while weighted scores account for class imbalance. F1 captures the balance between precision and recall and is the primary metric for slang detection.

5.3 Classification Models

5.3.1 Basic classification models

We experimented with two neural architectures, Convolutional Neural Networks (CNNs) (LeCun and Bengio, 1995) and Bidirectional Long Short-Term Memory (BiLSTM) networks (Liu and Guo, 2019; Pei et al., 2019), for slang classification. In a first setting, we trained CNN and BiLSTM classifiers using pre-trained word embeddings only, allowing us to assess the contribution of different embedding types in isolation. Specifically, we experimented with three types of pre-trained embeddings as input features: FastText, GloVe (Pennington et al., 2014), and BERT-based contextual embeddings (Devlin et al., 2019). CNNs capture local lexical patterns, while BiLSTMs model longer-range sequential dependencies within each textual instance. In a second setting, we followed the architecture proposed by (Pei et al., 2019) and evaluated CNN-CRF and BiLSTM-CRF models using the full feature configuration⁹ described in their work. This configuration combines word- and character-level representations with a set of linguistically motivated features, including part-of-speech (POS) tags, POS transition features, and pointwise mutual information (PMI) scores capturing atypical local word co-occurrence patterns associated with slang usage.

⁹Full features refers to the feature configuration proposed by Pei et al. (2019), which augments neural representations with linguistically motivated features such as part-of-speech (POS) information and pointwise mutual information (PMI) scores capturing atypical word co-occurrence patterns. We follow their implementation without modification.

5.3.2 Fine-tuning pre-trained transformer encoders

We fine-tuned four transformer models, BERT (Devlin et al., 2019), ALBERT (Lan, 2019), RoBERTa (Liu, 2019), and XLNet (Yang, 2019) using five-fold cross-validation over 30 epochs.¹⁰ Although trained under the same protocol, the models differ architecturally: BERT uses masked bidirectional language modelling; ALBERT reduces parameters via factorisation and weight sharing; RoBERTa extends BERT with training; and XLNet uses permutation-based modelling for bidirectional context.

5.3.3 Fine-tuning large language models (LLMs)

We fine-tuned several Large Language Models (LLMs) for slang classification, including GPT-4o¹¹ (version GPT-4o-2024-08-06), a high-capacity model, and GPT-4o-mini¹² (version GPT-4o-mini-2024-07-18), a smaller and more cost-efficient variant. We also fine-tuned LLaMA-3.1-8B-Instruct and LLaMA-3.1-70B-Instruct¹³, applying Low-Rank Adaptation (LoRA) to specialise these open-source models for slang detection. In addition to fine-tuning, we evaluated zero-shot (GPT-4o ZSp) and few-shot (GPT-4o FSp) prompting with structured task instructions.

5.4 Enhancing Slang Detection through Sentiment and Emotion Analysis

To capture affective cues in figurative slang, we incorporate automatically derived sentiment and emotion features into transformer-based and LLM models. Features were extracted from pre-trained emotion and sentiment classifiers¹⁵ and concatenated with contextual embeddings from ALBERT-xxlarge-v2 and BERT-large-uncased prior to classification. Although contextual language models encode affect implicitly, explicitly modelling sentiment and emotion appears to provide a targeted

¹⁰Each model was trained with three random seeds; performance variation was minimal (macro-F1 SD \approx 0.01).

¹¹<https://platform.openai.com/docs/models/gpt-4o>

¹²<https://platform.openai.com/docs/models/gpt-4o-mini>

¹³<https://huggingface.co/meta-llama/Llama-3.1-8B>

¹⁴Keras refers to embeddings created using https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding, which are randomly initialised and trained during the model’s learning process on task-specific data.

¹⁵Emotion embeddings from <https://huggingface.co/bhadresh-savani/bert-base-go-emotion>; sentiment embeddings from <https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>.

inductive bias for socially marked and evaluative slang usages, which may otherwise be diluted in general-purpose representations. The impact of affective features is shown in Table 4. Full prompts and parameter settings for all models are provided in Tables 14 and 13.¹⁶

6 Results

Deep learning architectures provided competitive baselines for slang detection. BiLSTM models using BERT embeddings achieved the strongest performance, with an accuracy of 0.85, outperforming variants trained on GloVe or FastText. This confirms that contextualised embeddings better capture slang-related semantic variation than static representations. CNN models performed reasonably well but struggled with longer or more polysemous constructions, reflecting a limited ability to model sequential dependencies. Incorporating POS and PMI features (Pei et al., 2019) did not improve performance. BiLSTM-CRF and CNN-CRF models underperformed simpler BiLSTM variants, particularly on the minority slang class.

Transformer-based models produced stronger results overall. BERT-large-uncased achieved the highest accuracy of 0.87 and an F1 score of 0.69 on slang instances. These results highlight the advantage of large bidirectional models for contextual sense disambiguation. RoBERTa and XLNet also performed well but showed slightly lower slang recall, suggesting difficulty in distinguishing figurative slang uses from literal ones in ambiguous contexts. This aligns with prior findings that BERT-style models capture subtle semantic distinctions more reliably (Sun et al., 2024).

Large language models showed a similar pattern. Fine-tuned GPT-4o-mini reached an accuracy of 0.86 and slightly outperformed GPT-4o despite its smaller size. This indicates that fine-tuning enhances sensitivity to slang-related cues and allows compact models to rival BERT-large-uncased. Zero-shot and few-shot GPT-4o performed well on non-slang usage but showed reduced recall for slang, suggesting that prompting alone does not fully resolve dual-meaning ambiguity. LLaMA-3.1 models underperformed relative to GPT-4o and BERT-based systems. The 70B variant produced strong non-slang performance, with a precision of 0.87 and a recall of 0.64, but struggled with slang

¹⁶This applies to all classification models described in this section.

(a) Deep learning models

Models	Features	Non-slang			Slang			Avg. Macro Scores			Avg. Weighted Scores			Acc
		Pr	Rec	F1	Pr	Rec	F1	Pr_M	Rec_M	F1_M	Pr_W	Rec_W	F1_W	
BiLSTM	GloVe	0.88	0.92	0.90	0.69	0.56	0.62	0.78	0.74	0.76	0.83	0.84	0.83	0.84
	BERT	0.86	0.96	0.91	0.78	0.47	0.58	0.82	0.72	0.75	0.84	0.85	0.83	0.85
	FastText	0.88	0.91	0.90	0.66	0.58	0.62	0.77	0.75	0.76	0.83	0.84	0.83	0.84
BiLSTM-CRF (full features) (Pei et al., 2019)	Keras ¹⁴	0.84	0.88	0.86	0.52	0.45	0.48	0.68	0.66	0.67	0.77	0.78	0.77	0.78
CNN	GloVe	0.87	0.87	0.87	0.61	0.61	0.61	0.74	0.74	0.74	0.81	0.81	0.81	0.81
	BERT	0.87	0.93	0.90	0.70	0.52	0.60	0.78	0.73	0.75	0.83	0.84	0.83	0.84
	FastText	0.88	0.91	0.90	0.66	0.58	0.62	0.77	0.75	0.76	0.83	0.84	0.83	0.84
CNN-CRF (full features) (Pei et al., 2019)	Keras	0.87	0.90	0.89	0.63	0.56	0.59	0.75	0.73	0.74	0.82	0.82	0.82	0.82

(b) Transformer and LLM models

Model	Non-slang			Slang			Avg. macro			Avg. weighted			Acc
	Pr	Rec	F1	Pr	Rec	F1	Pr_M	Rec_M	F1_M	Pr_W	Rec_W	F1_W	
BERT-L (Uncased)	0.90	0.94	0.92	0.76	0.63	0.69	0.83	0.79	0.80	0.86	0.87	0.87	0.87
RoBERTa-L	0.88	0.91	0.90	0.66	0.58	0.62	0.77	0.75	0.76	0.84	0.84	0.83	0.84
XLNet-L	0.87	0.95	0.90	0.60	0.49	0.54	0.73	0.72	0.72	0.83	0.82	0.82	0.82
ALBERT-XXL	0.89	0.94	0.92	0.76	0.61	0.68	0.82	0.78	0.80	0.86	0.86	0.85	0.86
GPT-4o (ZS)	0.94	0.79	0.86	0.54	0.82	0.65	0.74	0.80	0.75	0.80	0.80	0.80	0.80
GPT-4o (FS)	0.94	0.73	0.82	0.48	0.84	0.61	0.71	0.78	0.71	0.75	0.75	0.75	0.75
GPT-4o-Mini (FT)	0.91	0.92	0.91	0.72	0.66	0.69	0.81	0.79	0.80	0.86	0.86	0.86	0.86
GPT-4o (FT)	0.85	0.90	0.89	0.76	0.63	0.69	0.81	0.77	0.79	0.85	0.85	0.85	0.85
LLaMA-3.1-8B (FT)	0.37	0.62	0.47	0.86	0.62	0.72	0.61	0.62	0.59	0.75	0.67	0.53	0.67
LLaMA-3.1-70B (FT)	0.87	0.64	0.73	0.36	0.69	0.47	0.61	0.66	0.60	0.75	0.65	0.68	0.65
Models with Sentiment & Emotion													
BERT-L +S	0.90	0.95	0.92	0.79	0.64	0.76	0.84	0.80	0.84	0.88	0.88	0.88	0.88
BERT-L +E	0.90	0.94	0.92	0.78	0.63	0.76	0.84	0.79	0.84	0.87	0.87	0.87	0.87
BERT-L +S+E	0.90	0.95	0.92	0.81	0.65	0.78	0.86	0.80	0.85	0.89	0.89	0.89	0.89
ALBERT +S	0.90	0.93	0.92	0.77	0.66	0.74	0.83	0.80	0.83	0.87	0.87	0.87	0.87
ALBERT +E	0.90	0.92	0.91	0.76	0.66	0.73	0.83	0.79	0.82	0.87	0.87	0.86	0.87
ALBERT +S+E	0.91	0.92	0.91	0.72	0.69	0.77	0.82	0.81	0.84	0.88	0.88	0.88	0.88

Table 3: Model performance on the SlangTrack test set. Panel (a): deep learning models. Panel (b): transformer and large language models, including ablations with sentiment (S) and emotion (E) features. ZS = Zero-shot, FS = Few-shot, FT = Fine-tuned, S = Sentiment, E = Emotion.

detection, achieving a precision of 0.36 and a recall of 0.69. This indicates that without targeted fine-tuning, LLaMA models are less suited to the fine-grained distinctions required for slang disambiguation.

A per-word breakdown of BERT-large-uncased (Table 5) reveals substantial variation across target words. Items such as *climber* and *salty* achieve high macro-F1, while more ambiguous or infrequent slang senses (e.g., *cucumber*, *mammy*) yield lower slang-F1. This variation reflects differences in sense ambiguity and class balance, underscoring the need for per-word analysis beyond aggregate performance metrics.

Transformer fine-tuning produced the strongest performance, and adding sentiment and emotion features yielded small but consistent improvements. For BERT-large and ALBERT-xxlarge, affective cues increased macro-F1 by approximately 0.03–

0.05 and weighted F1 by around 0.01–0.02, primarily through slightly higher recall on slang instances. This suggests that affective information helps distinguish some figurative or evaluative uses from literal ones. Overall, affective features provide modest, model-specific gains and function as a supplementary rather than primary signal. Variance across runs is reported in Appendix 14. As shown in Table 10, improvements from sentiment and emotion features exceed the observed variance. Statistical significance is confirmed via paired bootstrap testing (Appendix 15).

7 Error Analyses

We analysed 100 misclassified instances from our top-performing slang detection model (BERT-L + Sentiment + Emotion), categorising them into types (Table 12).

Bad neighbours (23%): Misclassification often

Target Words	Example Sentences	Prediction Before Sentiment & Emotion	→	Sentiment	Emotion	Prediction After Sentiment & Emotion
Mammy	First 'black' woman to win an Oscar played a mammy slave. .. last one to win it also played a slave. Some see this as progress.	Non-Slang	→	Negative	Sadness	Slang
Germ	My room mate won't stop calling me a walking germ factory just because I have a cold.	Non-Slang	→	Negative	Anger	Slang
Chronic	North Korea is known for its chronic secrecy and isolation. But in recent years, despite its chronic struggles, there's been a noticeable rise in underground markets.	Slang	→	Neutral	Fear	Non-Slang
Eat	Just finished that brutal workout can't wait to eat! My abs are on fire, but hey, no pain, no gain, right?	Slang	→	Positive	Joy	Non-Slang

Table 4: Impact of sentiment and emotion on slang classification. The table shows examples of *Prediction before sentiment & emotion* and their corrected outputs using *Prediction after sentiment & emotion*.

Word	Cnt	F1 Non-slang	F1 Slang	Macro F1
Bmw	1,097	0.995	0.421	0.708
Brownie	964	0.872	0.800	0.836
Chronic	1,685	0.931	0.598	0.765
Climber	642	0.964	0.832	0.898
Cucumber	1,051	0.960	0.415	0.688
Eat	3,032	0.905	0.454	0.680
Germ	832	0.848	0.632	0.740
Mammy	1,070	0.929	0.458	0.693
Rodent	1,073	0.844	0.651	0.748
Salty	1,266	0.870	0.909	0.890

Table 5: Per-word F1 scores for BERT-large-uncased.

arises from the presence of nearby words that introduce strong pragmatic signals, such as abusive language, drug references, or harsh tones. Although such cues are not explicitly modelled in SlangTrack, they may co-occur with slang usage in natural data and implicitly influence model predictions. Importantly, SlangTrack does not aim to model toxicity or harmful intent; rather, these signals are treated as incidental contextual factors that can skew interpretation.

Proper nouns (10%): Proper nouns, especially those appearing as bi-grams or tri-grams, can confuse the model due to their compact and informal structure. These structures may be misinterpreted as slang or colloquial expressions, particularly when they lack distinguishing features.

Lost in length (13%): Very long instances with multiple clauses, or extremely short instances with limited contextual cues, can challenge the model. In such cases, important pragmatic signals may be diluted, truncated, or insufficiently represented in the input, reducing the model’s ability to distinguish slang from literal usage.

Polysemy (17%): Polysemy refers to target words with multiple meanings, both slang and non-slang, leading to misclassification when the immediate context is insufficient to disambiguate. For example, the word “germ” can mean a microorganism or

serve as a slang insult. Polysemy concerns the target word’s inherent multiple meanings rather than the surrounding context.

Ambiguity (7%): Ambiguity arises when the broader sentence context creates uncertainty, even if the target word’s meaning is clear in isolation. For instance, “salty” can mean “bitter” slang or “overly seasoned” literal. The word itself has a clear meaning in both cases, but without strong contextual cues, the model struggles to determine which sense is intended. Ambiguity is about the sentence creating uncertainty rather than the word having multiple inherent meanings.

Unknown (30%): Unconventional abbreviations, rare slang, or novel expressions pose challenges. Such terms often deviate from standard language patterns, making classification difficult, especially if they are absent from the model’s training data.

8 Conclusion

We address the task of slang detection by focusing on words that can be used in both slang and non-slang senses, an aspect that has received limited attention in prior work. This distinction enables a more fine-grained analysis of lexeme-specific and pragmatically driven meaning alternations, beyond aggregate classification accuracy. We introduce SlangTrack, a corpus of dual-meaning words, and benchmark a range of neural architectures, fine-tuned transformer-based language models, and large language models. Our results show that fine-tuned transformer models, particularly BERT-large-uncased, achieve the strongest performance on this task, while incorporating sentiment and emotion features yields modest but consistent improvements in ambiguous cases. SlangTrack thus provides a controlled benchmark for studying contextual slang usage across different lexical items.

9 Limitations

Our dataset integrates examples from heterogeneous sources that differ in register and historical period. CCOHA reflects earlier formal written English, while Twitter provides contemporary informal. These source differences may affect recall for slang senses that vary across time or context, but they also reflect real lexical change and support the study of sense variation across registers. Because CCOHA and Twitter differ substantially in corpus size and temporal span, we do not interpret raw frequency differences between them as direct evidence of diachronic slang trends; instead, our analyses focus on sense-level annotation and statistically controlled comparisons of slang versus non-slang usage across corpora. Although SlangTrack includes temporal metadata, the present study does not implement time-aware modelling, leaving explicit modelling of continuous semantic trajectories and slang adoption dynamics to future work.

While SlangTrack captures instance-internal pragmatic cues (e.g., affective stance and figurativity) expressed locally within individual textual instances, it does not explicitly model discourse-level pragmatics such as speaker alignment or conversational implicature spanning multiple interactions. The task therefore models local pragmatic interpretation rather than full discourse-level slang understanding. As with naturally occurring language data, slang and non-slang usages are not perfectly balanced across target words. This natural sparsity is addressed through macro-level evaluation and per-word analyses. Another limitation is the narrow lexical coverage (10 target words), adopted as a controlled pilot setting. Rather than maximising absolute performance, the results provide diagnostic insights into lexeme-dependent slang detection and motivate future work on broader coverage and richer contextual modelling.

References

Michael Adams. 2012. *Slang: The people’s poetry*. Oxford University Press.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte Im Walde. 2020. *CCOHA: Clean corpus of historical American English*. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6958–6966.

Michele Bevilacqua and Roberto Navigli. 2020. *Breaking through the 80% glass ceiling: Raising the state*

of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the conference-Association for Computational Linguistics. Meeting*, pages 2854–2864. Association for Computational Linguistics.

- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. 2016. *Demographic dialectal variation in social media: A case study of african-american english*. *arXiv preprint arXiv:1608.08868*.
- Julie Coleman. 2012. *The life of slang*. Oxford University Press, USA.
- Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. *Survey in characterization of semantic change*. *arXiv preprint arXiv:2402.19088*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Bert: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. *SlangNet: A WordNet like resource for English slang*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4329–4332.
- Bethany K Dumas and Jonathan Lighter. 1978. *Is slang a word for linguists?* *American Speech*, 53(1):5–17.
- Jacob Eisenstein. 2013. *What to do about bad language on the internet*. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369.
- Jonathon Green. 2015. *The vulgar tongue: Green’s history of slang*. Oxford University Press, USA.
- Anshita Gupta, Sanya Bathla Taneja, Garima Malik, Sonakshi Viji, Devendra K Tayal, and Amita Jain. 2019. *SLANGZY: A fuzzy logic-based algorithm for English slang meaning Selection*. *Progress in Artificial Intelligence*, 8:111–121.
- Janosch Haber and Massimo Poesio. 2024. *Polysemy—evidence from linguistics, behavioral science, and contextualized language models*. *Computational Linguistics*, 50(1):351–417.
- William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. *Diachronic word embeddings reveal statistical laws of semantic change*. *arXiv preprint arXiv:1605.09096*.
- Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. *GlossBERT: BERT for word sense disambiguation with gloss knowledge*. *arXiv preprint arXiv:1908.07245*.

- Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. [Slangvolution: A causal analysis of semantic change and frequency dynamics in slang](#). *arXiv preprint arXiv:2203.04651*.
- Chloe Kiddon and Yuriy Brun. 2011. [That’s what she said: double entendre identification](#). In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 89–94.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.
- Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. [Zero-shot word sense disambiguation using sense definition embeddings](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. [Diachronic word embeddings and semantic shifts: a survey](#). *arXiv preprint arXiv:1806.03537*.
- Z Lan. 2019. [Albert: A LiteBert for Self-supervised Learning of Language Representations](#). *arXiv preprint arXiv:1909.11942*.
- Yann LeCun and Yoshua Bengio. 1995. [Convolutional networks for images, speech, and time series](#). *The handbook of brain theory and neural networks*, 3361(10):1995.
- Gang Liu and Jiabao Guo. 2019. [Bidirectional LSTM with attention mechanism and convolutional layer for text classification](#). *Neurocomputing*, 337:325–338.
- Yinhan Liu. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Li Lucy and David Bamman. 2021. [Style variation and social meaning in online communities](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4745–4760. Association for Computational Linguistics.
- Theo Lynn, Patricia Takako Endo, Pierangelo Rosati, Ivanovitch Silva, Guto Leoni Santos, and Debbie Ging. 2019. [A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary](#). In *2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–8. IEEE.
- George A. Miller. 1995. [Wordnet: A lexical database for english](#). *Communications of the ACM*, 38(11):39–41.
- Ke Ni and William Yang Wang. 2017. [Learning to explain non-standard English words and phrases](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Alok Ranjan Pal and Diganta Saha. 2015. [Detection of slang words in e-data using semi-supervised learning](#). *arXiv preprint arXiv:1702.04241*.
- Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. [Slang detection and identification](#). In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 881–889.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. [Glove: Global vectors for word representation](#). In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. [Analyzing semantic change through lexical replacements](#). *arXiv preprint arXiv:2404.18570*.
- Rhys J Sandow, George Bailey, and Natalie Braber. 2024. [Language change is wicked: semantic and social meaning of a polysemous adjective](#). *English Language & Linguistics*, 28(1):135–156.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23. International Committee for Computational Linguistics.
- Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. [Toward informal language processing: Knowledge of slang in large language models](#). *arXiv preprint arXiv:2404.02323*.
- Zhewei Sun, Richard Zemel, and Yang Xu. 2021. [A computational framework for slang generation](#). *Transactions of the Association for Computational Linguistics*, 9:462–478.
- Zhewei Sun, Richard Zemel, and Yang Xu. 2022. [Semantically informed slang interpretation](#). *arXiv preprint arXiv:2205.00616*.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. [Survey of computational approaches to lexical semantic change detection](#). *Computational approaches to semantic change*, 6(1).
- Liang Wu, Fred Morstatter, and Huan Liu. 2018. [Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification](#). *Language Resources and Evaluation*, 52:839–852.
- Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. [Understanding slang with LLMs: Modelling](#)

cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.

Zhilin Yang. 2019. *Xlnet: Generalized autoregressive pretraining for language understanding*. *arXiv preprint arXiv:1906.08237*.

10 Implementation Details

Hardware Configuration

Deep Learning Models. All CNN and BiLSTM experiments were conducted on a single Tesla V100 GPU with 32 GB RAM using the Google Colab Pro+ platform.¹⁷

Fine-Tuning Experiments (LMs and LLMs). Fine-tuning for transformer-based language models and large language models (LLMs) was performed on a single NVIDIA A100 GPU (80 GB RAM).

Software Frameworks

All experiments used Python 3.10.12. Evaluation metrics were computed using scikit-learn.¹⁸ Transformer models were trained with three random seeds; performance variation was minimal (macro-F1 standard deviation 0.01).

Deep Learning Model Frameworks

TensorFlow¹⁹ and Keras²⁰ were used for implementing CNN and BiLSTM architectures. Hyperparameter tuning was performed using Keras Tuner.²¹ scikit-learn utilities were used to compute precision, recall, and F1-score.

Transformer Model Frameworks

SimpleTransformers²² and Hugging Face Transformers²³ enabled loading and fine-tuning of models such as BERT. Stratified K-fold cross-validation was implemented with scikit-learn.

Deep Learning (CNN and BiLSTM) Architecture Setup

Embedding Resources. FastText embeddings (wiki-news-300d-1M.vec), trained on Wikipedia

2017, UMBC WebBase, and statmt.org news (16B tokens), were used.²⁴ GloVe embeddings (300d, Common Crawl, 840B tokens) were loaded. BERT embeddings from bert-base-uncased were extracted for use in CNN and BiLSTM models.

Tokenisation and Sequence Processing. Sentences were tokenised using the Keras Tokenizer. Sequences were padded to a fixed length (95th percentile of training sentence lengths) using pad_sequences.

Hyperparameter Tuning. CNN and BiLSTM configurations were optimised via Keras Tuner’s Random Search. Fifty trials were run per architecture, tuning embedding dimensions, hidden sizes, dropout, learning rate, and optimiser. Selected hyperparameters appear in Table 13.

Fine-Tuning Procedures for GPT-4o and LLaMA Models

Training Setup and Specifications. GPT-4o, GPT-4o-mini, LLaMA-3.1-70B-Instruct-Reference, and LLaMA-3.1-8B-Instruct were fine-tuned using the same training and validation splits as BERT models. GPT-4o models were fine-tuned via OpenAI’s API. LLaMA models were fine-tuned using Together AI’s API,²⁵ incorporating LoRA for memory-efficient training. Each model was trained to classify sentences as slang or non-slang using structured JSONL prompt-completion pairs.

Structured Prompt Format. A consistent prompt template was used for GPT-4o and LLaMA models: Your task is to classify the sentence as either ‘slang’ or ‘non-slang.’ Please respond only with ‘slang’ or ‘non-slang.’

11 Exploratory ANOVA Analysis Between Slang and Non-Slang Texts

To provide a descriptive view of stylistic differences between slang and non-slang instances, we conducted an exploratory one-way ANOVA using automatically derived sentiment and emotion scores.²⁶

¹⁷<https://colab.research.google.com/>

¹⁸<https://scikit-learn.org/stable/>

¹⁹<https://www.tensorflow.org/>

²⁰<https://keras.io/>

²¹<https://keras-team.github.io/keras-tuner/>

²²<https://simpletransformers.ai>

²³<https://huggingface.co/transformers/>

²⁴<https://fasttext.cc/docs/en/english-vectors.html>

²⁵<https://together.ai>

²⁶All affective scores are produced by an external classifier and were not manually annotated; results should therefore be interpreted only as descriptive trends rather than causal linguistic evidence.

11.1 Sentiment Differences

Figure 1 shows the distribution of positive, neutral, and negative sentiment. ANOVA results (Table 6) indicate statistically detectable but modest differences between classes. Slang instances tend to be slightly more negative, while non-slang instances are more often neutral.

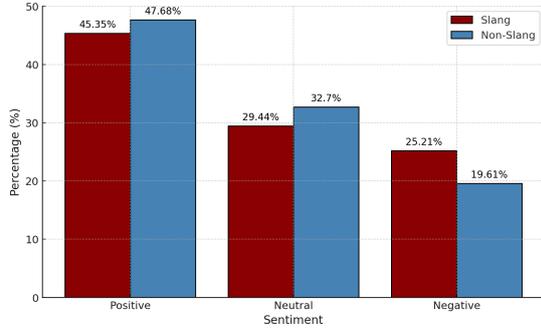


Figure 1: Sentiment distribution across slang and non-slang instances.

Sentiment	F	p	Significance
Positive	45.78	2.1×10^{-10}	Significant
Negative	38.54	4.3×10^{-8}	Significant
Neutral	12.92	3.0×10^{-4}	Significant

Table 6: Exploratory ANOVA results for sentiment.

11.2 Emotion Differences

Emotion scores (Figure 2) show detectable differences for a subset of categories (Table 7). Slang instances exhibit slightly higher anger and sadness, while other emotions show no meaningful differences.

Emotion	F	p	Significance
Anger	9.61	1.9×10^{-3}	Significant
Sadness	5.35	3.7×10^{-2}	Significant
Fear	0.11	0.74	Not Significant
Joy	0.32	0.57	Not Significant
Love	1.52	0.22	Not Significant
Surprise	0.03	0.85	Not Significant

Table 7: Exploratory ANOVA results for emotion categories.

12 Appendix: Full Sense Inventories for SlangTrack Target Words

Table 8 lists all attested senses for each target word in the SlangTrack dataset. Sense IDs (S1–Sn) correspond to the annotation scheme used throughout

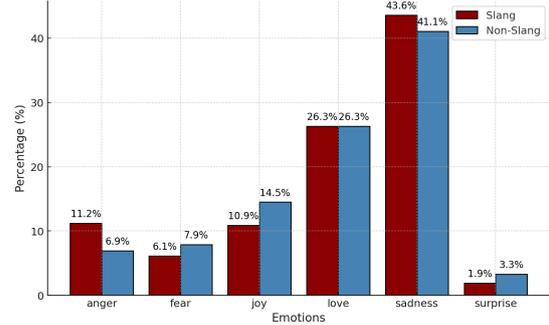


Figure 2: Emotion score distribution across slang and non-slang instances.

the paper. This table enumerates the sense inventories only.

13 Statistical Analysis of Diachronic Shifts in Slang vs. non-slang Sense Usage

We evaluate whether the relative frequency of pre-annotated slang versus non-slang senses differs between COHA (historical data) and Twitter (contemporary data). Each annotated token represents a binary outcome (slang or non-slang), enabling comparison of proportions across periods using the standard two-proportion z -test. This analysis examines redistribution in the usage of existing senses rather than the emergence of new ones.

13.1 Two-Proportion z -Test

For a given word w , let $p_{\text{COHA}}(w)$ and $p_{\text{Twitter}}(w)$ denote the observed proportions of slang-labelled tokens in COHA and Twitter. The null hypothesis states that the true slang probabilities are equal:

$$H_0 : p_{\text{COHA}}(w) = p_{\text{Twitter}}(w),$$

$$H_1 : p_{\text{COHA}}(w) \neq p_{\text{Twitter}}(w).$$

The pooled estimate under H_0 is:

$$\hat{p} = \frac{s_{\text{COHA}} + s_{\text{Twitter}}}{n_{\text{COHA}} + n_{\text{Twitter}}},$$

where s_{COHA} and s_{Twitter} are slang counts, and n_{COHA} , n_{Twitter} are total token counts in each corpus. The standard error of the difference is:

$$SE = \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_{\text{COHA}}} + \frac{1}{n_{\text{Twitter}}} \right)}.$$

The test statistic is:

$$z = \frac{\hat{p}_{\text{Twitter}} - \hat{p}_{\text{COHA}}}{SE}.$$

Word	Sense Inventory (S1–Sn)
Eat	S1: Consume food. S2: Perform oral sex on a woman. S3: Make money or absorb a financial loss (“take the loss”). S4: Rob someone in a low-violence street context. S5: Defeat, destroy, or overwhelm. S6: Irritate or annoy.
BMW	S1: German automobile brand. S2: Derogatory reference to Black individuals (“Black Man Working,” etc.). S3: Acronym for “Be My Wife.”
Brownie	S1: Chocolate dessert. S2: Silly or foolish person. S3: Racially offensive reference to brown-skinned individuals. S4: Marijuana-infused edible. S5: Vulgar reference to the anus. S6: Unit of social credit (“brownie points”). S7: Personal name or nickname. S8: Frequent collocate used in fixed expressions.
Chronic	S1: High-quality cannabis. S2: Medical term for a long-lasting or persistent condition. S3: Excellent or outstanding. S4: Negative extreme (e.g., a “chronic” habit or person). S5: Proper name (album, song, title).
Climber	S1: Person attempting upward social mobility. S2: Rock climber (sport). S3: Climbing plant. S4: Burglar who gains access by climbing.
Germ	S1: Microorganism. S2: Cigarette slang used in institutional settings. S3: Offensive term for people of German descent. S4: Contemptible or unpleasant person. S5: Proper name (company or brand). S6: Frequent collocate in compounds.
Mammy	S1: Racialised stereotype of a Black woman domestic worker. S2: Dialectal or literal “mother.” S3: Slang for a large amount (e.g., “money’s mammy”). S4: Proper name (brand, company, or artistic title).
Cucumber	S1: Penis (slang). S2: Vegetable. S3: Proper name (company, brand, or software).
Rodent	S1: Insult for someone unattractive, untrustworthy, or unintelligent. S2: Animal of the order <i>Rodentia</i> .
Salty	S1: Irritated, annoyed, or resentful. S2: Tasting of salt. S3: Old, worn, or well-used. S4: Tough, aggressive, or hardened. S5: Crude, obscene, or vulgar. S6: Proper name (animal, brand, or song title).

Table 8: Complete sense inventories for all SlangTrack target words. Slang and non-slang senses are shown together. Sense IDs (S1–Sn) correspond to those used in the annotation guidelines.

Two-sided p -values are obtained from the standard normal distribution.

13.2 Worked Example: *Mammy*

For *Mammy*, the annotated sense inventory includes both slang and non-slang meanings. Slang senses: (1) stereotyped Black woman (“mammy” caricature), (3) figurative abundance sense (e.g., “money’s mammy”). Non-slang senses: (2) literal ‘mother’, (4) proper-name, company, or song title use.

COHA. Slang tokens: 97. Total tokens: 194.

$$\hat{p}_{\text{COHA}} = 97/194 = 0.500.$$

Twitter. Slang tokens: 69. Total tokens: 876.

$$\hat{p}_{\text{Twitter}} = 69/876 \approx 0.0788.$$

Pooled estimate.

$$\hat{p} = \frac{97 + 69}{194 + 876} = \frac{166}{1070} \approx 0.1551.$$

Standard error.

$$SE \approx 0.0287.$$

Test statistic.

$$z \approx -14.66, \quad p < 0.0001.$$

This indicates a substantial decrease in the slang usage of *Mammy* from COHA to Twitter, with the racialised sense becoming rare in contemporary data.

13.3 Results for All Target Words

Table 9 reports slang proportions for COHA and Twitter, their differences, and z -test significance levels.

13.4 Interpretation

Three items show statistically significant redistribution in slang usage between COHA and Twitter: *Salty* (increase), *Mammy* (decrease), and *Chronic* (decrease). For all other words, changes are small

Word	p_{COHA}	$p_{Twitter}$	Δp	z	p	Sig.
Eat	0.100	0.120	+0.020	1.67	0.0960	-
BMW	0.022	0.012	-0.010	-1.16	0.2450	-
Brownie	0.283	0.265	-0.018	-0.45	0.6550	-
Chronic	0.280	0.225	-0.055	-2.59	0.0097	**
Climber	0.193	0.219	+0.026	0.65	0.5150	-
Germ	0.110	0.091	-0.019	-0.78	0.4370	-
Mammy	0.500	0.079	-0.421	-14.66	<0.0001	***
Cucumber	0.057	0.072	+0.015	0.69	0.4890	-
Rodent	0.292	0.309	+0.017	0.46	0.6470	-
Salty	0.193	0.742	+0.549	18.12	<0.0001	***

Table 9: Slang usage proportions for COHA vs. Twitter. Significance: * $p < .01$, ** $p < .001$ (two-sided).

and not significant. The results illustrate diachronic variation in the distribution of attested senses, consistent with usage-based accounts of sense competition and shifting pragmatic salience.

14 Variance Estimation

To provide a transparent account of performance variability, we report macro-F1 scores together with variance estimates for all models (see Table 10). Transformer models were fine-tuned using three distinct random seeds, and we therefore report macro-F1 \pm standard deviation across these runs. For deep learning models, which are deterministic under our setup, variance was estimated using 1,000-sample non-parametric bootstrap resampling over the test predictions.

Model	Non-slang F1	Slang F1	Macro-F1 \pm SD
Deep Learning Models			
BiLSTM (BERT)	0.91	0.58	0.75 \pm 0.012
BiLSTM (GloVe)	0.90	0.62	0.76 \pm 0.015
BiLSTM (FastText)	0.90	0.62	0.76 \pm 0.014
CNN (BERT)	0.90	0.60	0.75 \pm 0.017
CNN (GloVe)	0.87	0.61	0.74 \pm 0.013
CNN (FastText)	0.90	0.62	0.76 \pm 0.012
BiLSTM-CRF (full features)	0.86	0.48	0.67 \pm 0.020
CNN-CRF (full features)	0.89	0.59	0.74 \pm 0.019
Transformer Models			
BERT-large-uncased	0.92	0.69	0.80 \pm 0.010
RoBERTa-large	0.90	0.62	0.76 \pm 0.011
XLNet-large	0.90	0.54	0.72 \pm 0.012
ALBERT-xxlarge-v2	0.92	0.68	0.80 \pm 0.010
Models with Sentiment & Emotion			
BERT-L +S	0.92	0.76	0.84 \pm 0.010
BERT-L +E	0.92	0.76	0.84 \pm 0.011
BERT-L +S+E	0.92	0.78	0.85 \pm 0.012
ALBERT +S	0.92	0.74	0.83 \pm 0.011
ALBERT +E	0.91	0.73	0.82 \pm 0.012
ALBERT +S+E	0.91	0.77	0.84 \pm 0.013

Table 10: Performance of deep learning and transformer models on the SlangTrack test set, including models enriched with sentiment (S) and emotion (E) features.

15 Statistical Significance of Sentiment and Emotion Features

To verify that the performance improvements from sentiment and emotion features are statistically meaningful, we conducted paired significance test-

Model	Comparison	Macro-F1	p -value
BERT-large-uncased	Base \rightarrow +S	+0.04	0.012
	Base \rightarrow +E	+0.04	0.018
	Base \rightarrow +S+E	+0.05	0.003
ALBERT-xxlarge-v2	Base \rightarrow +S	+0.03	0.021
	Base \rightarrow +E	+0.02	0.025
	Base \rightarrow +S+E	+0.04	0.009

Table 11: Paired bootstrap significance testing for sentiment and emotion features. Each block shows the comparisons between the base model and variants with sentiment (S), emotion (E), or both.

ing on the transformer models.

Test procedure. For each transformer-based model, we compare the base fine-tuned model with its affect-enhanced variants using paired bootstrap resampling (Koehn, 2004). Resampling is performed over test predictions from a fixed fine-tuning configuration, using identical test instances for both models. We draw 10,000 bootstrap samples with replacement and compute the difference in macro-F1 for each sample, yielding an empirical distribution of performance differences. Two-sided p -values are estimated as the proportion of samples in which the difference includes zero.

Results. Table 11 reports the mean macro-F1 improvements (Δ Macro-F1) and associated p -values. For both BERT-large and ALBERT-xxlarge, the combined sentiment and emotion configuration (+S+E) yields statistically significant improvements ($p < 0.05$). Individual sentiment-only (+S) and emotion-only (+E) features produce smaller gains that are not consistently significant across models. These findings are consistent with the low run-to-run variance observed across random seeds (Appendix 14), indicating that the improvements are robust to sampling variation rather than artefacts of random initialisation.

16 Annotation Challenges in Historical Corpus Data

These examples demonstrate why annotating words that have both slang and non-slang senses is cognitively demanding, particularly in long, descriptive, historical corpora such as COHA. One of our target words, *salty*, can express a wide range of meanings, and the annotator must determine which sense is intended based on extended narrative context. These examples illustrate how COHA’s long, descriptive, and multi-layered passages often involving figurative language, shifting narrative perspectives, and domain-specific idioms significantly increase the

complexity of sense annotation. As a result, annotations drawn from historical corpora require more time, attention, and contextual reasoning than annotations drawn from short, contemporary, or informal texts.

Sense Inventory for *salty*

- **S1: Irritated, annoyed, or resentful** slang sense referring to emotional upset or reactive behaviour.
- **S2: Tasting of salt** — literal reference to flavour, salinity, or physical salt content.
- **S3: Old, worn, or well-used** describes objects or individuals that appear aged or weathered.
- **S4: Tough, aggressive, or hardened** slang sense used for rugged, experienced, or hardened individuals.
- **S5: Crude, obscene, or vulgar** slang sense used for coarse or inappropriate language or behaviour.
- **S6: Proper name** used when *salty* appears as a title or name (animal, brand, song title).

Example 1 (COHA T1)

My feelings are hurt, and no one loves me. It doesn't matter what I look like. A large bowl of snack food or a box of chocolates is so 'good' that it disappears. (e) "I am just as good, strong, and brave as they are." If I eat heart, kidneys, and liver, then I will be strong like a lion. (f) "I almost starved to death during period X." People who have been deprived of food continuously often will eat whatever they can, including rich and salty snack foods @ @ @ @ @ @ @ @ @ @ d. Due to the following aspects: (a) to make it less extreme, when a person who habitually consumes a high-fat diet must reduce fat intake to a moderate level; (b) to limit or restrict the meaning of and suggests a difference that limits, restricts, or adapts to a new purpose, as when one who has habitually consumed a high-fat, rich diet must follow a low-fat, low-calorie diet until sufficient weight is lost that the gall bladder can be removed; (c) to make a minor change in, as a patient with an ulcer. Such dietary adjustments are often recommended gradually rather than abruptly. Medical guidance typically emphasizes moderation rather

than total elimination. Patients are advised to monitor portion size and ingredient content carefully. Attention to fat and salt intake is presented as part of long-term health management.

Sense-Identification Notes:

This passage blends nutritional explanation, emotional examples, and medical reasoning, making the annotator process several shifts in meaning before encountering the literal use of salty. The long, layered structure increases the cognitive load and requires careful reading. The intended sense is a literal reference to salty food.

Label: S2 Non-slang

Example 2 (COHA T1)

Back into the hole you crawled @ @ @ @ @ @ @ @ @ @ and get out of here. Hey, hey, hey, hey, hey. Hey, what're you doing? Pretty hot out here. Must be what, 90? It's gonna get real toasty in there. Hey, now, now, let's not get salty here — no need to get angry. I'll tell you what. You give everything you got in the register, forget about pumping gas, and we'll be on our merry way. The purse, too. Please, go away now. No! No, don't! No! No, don't! (SHOUTING) Whoa, whoa! Could you hold it there? MAN: Ging, guy. Baker, what is this? That's, um, caviar. You know, that stuff comes all the @ @ @ @ @ @ @ @ @ @ — you kidding me? I get a stupid little sea wrapped pound cake with a candle on it for my birthday, and this guy's spending \$20 an ounce on this stuff? If you had the stars, it would be a whole different war. Yeah. How do you eat it? Please put it on one of those crackers, Danny don't get salty about it. RUIZ: Oh, that's disgusting. Danny, Danny, man, I got to clean that stuff up... Hey, keep your voice down, all right? You don't need to snap at everyone in the room. Everybody's already on edge, and you're making it worse. Take a breath and calm yourself down. This doesn't have to turn ugly if you don't let it. Now move, before someone really loses their temper.

Sense-Identification Notes:

The expression "let's not get salty here" appears during an escalating confrontation marked by shouting, threats, and attempts to control anger. In this context, salty clearly refers to emotional irritation or annoyance rather than taste or physical properties. The surrounding dialogue explicitly references anger management ("no need to get angry," "calm yourself down," "loses their temper"), which firmly supports interpretation as S1.

Label: S1 Slang

Example 3 (COHA T2)

Third Son ends up with the princess and half the kingdom. He fitted right in. Only Emily knew he didn't belong, and it gave a kind of edge to his performance, she felt. She ran him through his lines herself. (Leon played the older two sons.) She put an extra, salty twang in his voice. The real Third Son, meanwhile more handsome, with less character – lay face-up backstage, grinning vacantly. Emily had never actually planned to be @ @ @ @ @ @ @ @ @ @, thought of it as temporary work. "They would taste now; it's hot in here," she told him. @ @ @ @ @ @ @ @ @ @ his jaw, a tiny muscle into which he poured all the concentrated tension of his body. "You invited me here, lady..." He stood straighter, setting his shoulders as if bracing for impact. The words came out sharper than before, clipped and deliberate. Emily nodded, encouraging him to hold that tone. The edge made the character sound older, tougher, less naïve. It was no longer a boy pleading, but a man demanding. Even from backstage, the difference in force was audible. The performance carried a harder, more aggressive energy. The added roughness gave the scene its weight.

Sense-Identification Notes:

In this theatrical context, salty modifies twang to describe a deliberately sharpened and hardened vocal quality. The surrounding performance-related cues (e.g., tension, force, and aggression in delivery) indicate a stylistic choice rather than a literal taste or emotional irritation. This usage aligns with the slang sense denoting toughness or hardness.

Label: S4 Slang

Error Category	Examples	Reason for Misclassification	Gold Label → Predicted Label
Bad Neighbors	I think y’all understand the intense hate and fear for that rodent -looking motherfucker.	The strong slang word “motherfucker” triggers misclassification, while “rodent” adds a negative tone but isn’t slang.	Non-Slang → Slang
Proper Nouns	Wow, believe still remember brownie smile song girl scout memories. Good burger, man. I wish you could come to the sweet brownie party.	Phrases like “Brownie Smile Song” and “sweet brownie party” are proper nouns. Informal phrasing misleads the model into treating them as slang.	Non-Slang → Slang
Lost in Length	Post-1960s growth, a small, expensive underclass resulted in structural problems... chronic joblessness and welfare dependency.	Long, complex sentences make identifying context difficult, and pre-processing can reduce clarity.	Non-Slang → Slang
Polysemy	@Officer_Grayson Once a germ , always a germ . He’s as unclean as pork.	The word “germ” has multiple meanings, either as a microorganism or an insult. Lack of clear context causes errors.	Non-Slang → Slang
Polysemy	As a tiny rodent .. I see things from a unique angle. Like that guy over there... he’s not wearing under-pants.	The metaphor “tiny rodent” was interpreted literally instead of as slang, leading to misclassification.	Slang → Non-Slang
Ambiguity	The book’s protagonist is a mammy figure who is both nurturing and deeply flawed, and becomes a symbol of resistance against systemic oppression.	The context links “mammy” to literary analysis, suggesting non-slang usage, but informal or stereotypical connotations mislead the model.	Non-Slang → Slang
Ambiguity	My mom is really starting to get on my fucken nerves being the germ freak she is.	The word “germ” can be literal (bacteria) or slang (obsession with cleanliness). Ambiguous context caused misclassification.	Slang → Non-Slang
Unknown	Ugh, can’t wait to eat something after this workout! Abs are killing me, lol hoebag move, though.	Informal abbreviations like “lol” and rare slang terms like “hoebag” confuse the model.	Non-Slang → Slang

Table 12: Examples of misclassified samples for each error category.

Deep Learning, Transformer, and Large Language Models	
Model	Parameters
Fine-Tuned Transformer Models (BERT, RoBERTa, ALBERT, XLNet)	num train epochs = 30, learning rate = 4e-5, train batch size = 64, eval batch size = 64, Optimiser = AdamW
Fine-Tuned BERT and ALBERT with Sentiment and Emotion Analysis	num train epochs = 30, learning rate = 4e-5, train batch size = 64, eval batch size = 64, Optimiser = AdamW
BiLSTM + GloVe embeddings	Embedding Dimension = 300, BiLSTM Units = 32, Dense Units = 64, Dropout Rate = 0.2, Optimiser = Adam, Learning Rate = 2.93e-03, epochs = 30
BiLSTM + FastText embeddings	Embedding Dimension = 300, BiLSTM Units = 256, Dense Units = 64, Dropout Rate = 0.3, Optimiser = Adam, Learning Rate = 7.02e-04, epochs = 30
BiLSTM + BERT embeddings	Embedding Dimension = 768, BiLSTM Units = 128, Dense Units = 128, Dropout Rate = 0.3, Optimiser = RMSprop, Learning Rate = 7.44e-03, epochs = 30
CNN + GloVe embeddings	Embedding Dimension = 300, Conv Units = 128, Dense Units = 32, Dropout Rate = 0.4, Optimiser = Adam, Learning Rate = 1.00e-03, epochs = 30
CNN + FastText embeddings	Embedding Dimension = 300, Conv Units = 224, Dense Units = 128, Dropout Rate = 0.2, Optimiser = RMSprop, Learning Rate = 1e-03, epochs = 30
CNN + BERT embeddings	Embedding Dimension = 768, Conv Units = 64, Dense Units = 32, Dropout Rate = 0.3, Optimiser = RMSprop, Learning Rate = 1e-03, epochs = 30
Fine-Tuned LLaMA Models (LLaMA-3.1-70B-Instruct, LLaMA-3.1-8B-Instruct)	num train epochs = 30, learning rate = 1e-5, train batch size = 64, LoRA Enabled = True, LoRA Rank = 64, LoRA Alpha = 128, LoRA Dropout = 0.0, LoRA Trainable Modules = all-linear, Optimiser = AdamW, Learning Rate Scheduler = Linear, Max Grad Norm = 1.0
Fine-Tuned OpenAI Models (GPT-4o, GPT-4o-mini)	num train epochs = 3, batch size = 20, learning rate = 1e-5, Optimiser = AdamW, Temperature = 0, Max Tokens = 1024, Top p = 1.0, Frequency Penalty = 0.0, Presence Penalty = 0.0
GPT-4o ZS and GPT-4o FS	Max Tokens = 2048, Temperature = 0, Top p = 0.9, Frequency Penalty = 0.0, Presence Penalty = 0.0
Machine Learning Models	
Estimator	Hyperparameters
Logistic Regression (LR)	penalty=l2, C=1.0, solver=lbfgs, max_iter=100, verbose=0
Support Vector Machine (SVM)	C=1.0, gamma=1.0, cache_size=200, max_iter=-1
Random Forest (RF)	n_estimators=100, max_depth=10, min_samples_split=2
AdaBoost	n_estimators=50, learning_rate=1.0, base_estimator=DecisionTreeClassifier, algorithm=SAMME.R
CatBoost	iterations=1000, learning_rate=0.03, depth=6, verbose=True

Table 13: Parameters for Deep Learning, Transformer-Based, and Machine Learning Models. The first section presents trainable model parameters, while the second section highlights hyperparameters for machine learning models.

Strategy	Prompt	Example
Zero-shot prompting	<p>Task explanation</p> <p>Explicit behavioural guidelines</p> <p>Task instance</p>	<p>Your task: Classify a given sentence as either <i>slang</i> or <i>non-slang</i>.</p> <ol style="list-style-type: none"> FIRST line: ONLY write <i>slang</i> or <i>non-slang</i> with no extra words or punctuation. Following lines: Explain WHY the sentence is classified as slang or non-slang. Ensure the response follows this format for accurate extraction. <p>Task instance: <i>Ravens fans, players still salty, got ass kicked last week.</i></p> <p>Expected Answer: Slang</p> <p>Reasoning: The phrase <i>still salty</i> is slang for being upset or bitter about a past event. The phrase <i>got ass kicked</i> further emphasises informality.</p>
Few-shot prompting	<p>Task explanation</p> <p>Explicit behavioural guidelines</p> <p>Example instances</p> <p>Task instance</p>	<p>Your task: Classify a given sentence as <i>slang</i> or <i>non-slang</i>.</p> <ol style="list-style-type: none"> FIRST line: ONLY write <i>slang</i> or <i>non-slang</i> with no extra words or punctuation. Following lines: Explain WHY the sentence is classified as slang or non-slang. Ensure the response follows this format for accurate reasoning extraction. <p>Example instance: <i>Man, that chronic had me feeling way too relaxed last night.</i> Answer: Slang</p> <p>Reasoning: The term <i>chronic</i> is slang for high-quality marijuana.</p> <p>Example instance: <i>Scientists discovered a new type of germ in the petri dish.</i> Answer: Non-slang</p> <p>Reasoning: The word <i>germ</i> is used literally in a microbiological context.</p> <p>Task instance: <i>Ravens fans, players still salty, got ass kicked last week.</i> Expected Answer: Slang</p> <p>Reasoning: The phrase <i>still salty</i> expresses resentment, and <i>got ass kicked</i> adds to the informal slang tone.</p>

Table 14: Zero-shot and Few-shot prompting strategies for slang detection.

Statistical Semantic Change Detection via Usage Similarities

Taichi Aida

Hitotsubashi University
taichia@scl.sds.hit-u.ac.jp

Daichi Mochihashi

The Institute of Statistical Mathematics
daichi@ism.ac.jp

Hiroya Takamura

Artificial Intelligence
Research Center, AIST
takamura.hiroya@aist.go.jp

Toshinobu Ogiso

National Institute for Japanese
Language and Linguistics
togiso@ninjal.ac.jp

Mamoru Komachi

Graduate School of SDS
Hitotsubashi University
mamoru.komachi@r.hit-u.ac.jp

Abstract

Semantic change detection comprises two sub-tasks: **classification**, which predicts whether a target word has undergone a semantic shift, and **ranking**, which orders words according to the degree of their semantic change. While most prior studies concentrated on **ranking** subtask, the **classification** subtask plays an equally important role, since many practical scenarios require a yes/no decision on semantic change rather than a global ranking. In this work, we propose a novel statistical method that predicts the presence or absence of semantic change. While most existing approaches infer semantic change by comparing word embeddings across time periods or domains, our method directly models the diachronic/synchronic consistency of usage-level similarity scores. Our experiments on SemEval-2020 Task 1 and WUGS datasets demonstrate that the proposed formulation outperforms existing state-of-the-art embedding-based methods, and robustly detects semantic change across languages in both diachronic and synchronic settings.¹

1 Introduction

The meanings of words naturally evolve over time and across domains. Detecting such semantic change is essential for linguistic and lexicographic research, as well as for studying cultural and societal dynamics (Traugott and Dasher, 2001; Cook and Stevenson, 2010). Beyond these humanities-oriented applications, recent work has highlighted the importance of semantic change detection for various additional purposes, including information retrieval (Kutuzov et al., 2018) and efficient updating of masked language models (Su et al., 2022).

The Semantic Change Detection (SCD) task aims to automatically identify words that have undergone semantic shift. Recent shared tasks

¹Source code is available at <https://github.com/a1da4/usage-similarity-polya>.

Method	EN	DE	LA	SV
SGNS (Rother et al., 2020)	73.0	54.2	45.0	61.3
SGNS (Pražák et al., 2020)	62.2	75.0	70.0	67.7
BERT (Asgari et al., 2020)	70.3	75.0	55.0	74.2
Pólya (ours)	76.1	80.0	N/A	88.6

Table 1: Accuracy (in %) in SemEval-2020 Task 1 (Schlechtweg et al., 2020); our method does not rely on word embeddings contrary to prior state-of-the-arts.

such as SemEval-2020 Task 1 (Schlechtweg et al., 2020) and the WUGS (Schlechtweg et al., 2021) framework have established benchmark datasets and evaluation protocols for this task. There are two subtasks in SCD: **classification**, which predicts whether a target word is semantically changed, and **ranking**, which orders target words according to the degree of semantic change (Schlechtweg et al., 2020). Due to the unsupervised nature of the problem, most prior studies have focused on the ranking subtask, evaluating models based on similarity scores derived from word embeddings across two periods/domains (Rosin et al., 2022; Rosin and Radinsky, 2022; Cassotti et al., 2023; Periti and Tahmasebi, 2024; Aida and Bollegala, 2024). However, ranking-based evaluation has inherent limitations, including limited interpretability of raw similarity scores and uncertainty about which parts of the ranked list are reliable. These issues motivate a stronger emphasis on the classification subtask.

In this work, we propose a new method that statistically determines whether a word has undergone semantic change by assessing the diachronic/synchronic consistency of a set of usage-level similarity scores. We consider SCD under the assumption that usage-level similarity scores are available, as is the case in current benchmark datasets such as SemEval-2020 Task 1 and WUGS, where these scores are provided via human annotations. Rather than addressing how such similarity scores should be predicted, our focus is on how semantic change can be statistically determined once

usage-level similarity information is given. This design choice allows us to study SCD independently of specific similarity estimation methods. To this end, our framework models these similarity distributions using the Pólya distribution, enabling us to test whether the scores from two time periods are likely to originate from the same underlying distribution (indicating semantic stability) or from distinct distributions (indicating semantic change).

2 Related Work

Both classification and ranking evaluations in the SemEval-2020 Task 1 (Schlechtweg et al., 2020) are derived from WUGS-style annotation graphs (Schlechtweg et al., 2021), which are constructed by collecting a fixed number of usage examples from two different time periods/domains for each target word. After that, human annotators rate pairs of usages on a four-point semantic similarity scale (Schlechtweg et al., 2018, 2024b), and the weighted graph is processed to obtain both binary labels and continuous scores representing its degree of semantic change. These resources have provided a baseline for evaluating SCD systems.

Methods for SCD generally rely on comparing static (Kim et al., 2014; Kulkarni et al., 2015; Yao et al., 2018; Aida et al., 2021) or contextualized (Hu et al., 2019; Giulianelli et al., 2020; Rosin et al., 2022; Rosin and Radinsky, 2022; Cassotti et al., 2023; Aida and Bollegala, 2024) word embeddings. Despite these advances, existing methods focus on the ranking subtask, in part because similarity-based metrics naturally yield ordered scores. Therefore, open fundamental challenges in binary decision-making still remain.

3 Method

In this work, we also leverage the WUGS-style dataset described above. For a given word (e.g., *plane*), we collect N usages from texts in period A (e.g., ‘*in the horizontal plane*’) and M usages from period B (e.g., ‘*the plane was in flight*’).

We construct a matrix \mathbf{X} of size $(N+M) \times (N+M)$ over all possible pairs of these $(N+M)$ usages, as shown in Figure 1, where some of its entries x_{ij} are given a similarity score 1 through 4 between usage i and usage j . We consider two cases with this matrix:

Without sense change. In this case, there is no distinction between the entries x_{ij} in \mathbf{X} and thus all the annotated scores could be assumed to be

$$\mathbf{X} = \begin{matrix} & & & 1 & & N & N+1 & & N+M \\ & & & 1 & & & & & \\ & & & 4 & & & & & 2 \\ & & & & 4 & 3 & & & & 3 & 4 \\ & & & & & 2 & & & & 2 & \\ & & N & & 1 & & & & & 2 & \\ & & N+1 & & 1 & & 4 & & & 4 & 1 \\ & & & & & & 2 & & & & \\ & & & & & & & 2 & & 3 & \\ & & & & & & & & & & 4 \\ & & & & & & & & & & N+M \end{matrix}$$

Figure 1: Example of the score matrix \mathbf{X} of usage similarities. Some of the usage pairs are annotated to have similarity scores, 1–4 in this case.

generated from the same underlying distribution $\mathbf{p} = (p_1, p_2, p_3, p_4)$ over the scores 1 through 4:

$$x_{ij} \sim \mathbf{p}_0 \text{ i.i.d. } (1 \leq i, j \leq N+M) \quad (1)$$

With sense change. If there is a semantic change between the periods A and B, the similarity score over the pairs of usages within the same period or between the different periods will differ. Therefore, entry x_{ij} is assumed to be generated from one of the four distinct distributions \mathbf{p}_n ($n = 1, \dots, 4$) for the associated block of \mathbf{X} delineated in Figure 1.

$$x_{ij} \sim \begin{cases} \mathbf{p}_1 & (1 \leq i, j \leq N) \\ \mathbf{p}_2 & (1 \leq i \leq N, N+1 \leq j \leq N+M) \\ \mathbf{p}_3 & (N+1 \leq i \leq N+M, 1 \leq j \leq N) \\ \mathbf{p}_4 & (N+1 \leq i, j \leq N+M) \end{cases} \quad (2)$$

As a generative model we need a prior for $\mathbf{p} = \mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_4$, and the simplest choice is a Dirichlet distribution

$$p(\mathbf{p}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1} \quad (3)$$

where $K = 4$ in our case and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ is a hyperparameter for this prior distribution. We employed $\boldsymbol{\alpha} = (1, \dots, 1)$, i.e., uniform distribution over probability simplex, throughout this study.

Let a binary latent variable θ denote whether there is no semantic change ($\theta = 0$) or there is a change ($\theta = 1$). Then the likelihood of the data \mathbf{X} when $\theta = 0$ is given as follows:

$$\begin{aligned} p(\mathbf{X}|\theta=0) &= p(\mathbf{X}) = \int p(\mathbf{X}|\mathbf{p})p(\mathbf{p})d\mathbf{p} \\ &= \int \prod_{i,j} \prod_{k=1}^K p_k^{\mathbb{I}(x_{ij}=k)} \cdot \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1} d\mathbf{p} \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(L + \sum_k \alpha_k)} \prod_{k=1}^K \frac{\Gamma(\alpha_k + n_k)}{\Gamma(\alpha_k)} \end{aligned} \quad (4)$$

where $n_k = \sum_{ij} \mathbb{I}(x_{ij} = k)$ is the frequency of score k in \mathbf{X} and L is the number of annotated entries.

This formula (4) is known as a Pólya distribution (Minka, 2000; Murphy, 2022). For the case semantic change, the likelihood is a product of block-wise Pólya distributions according to the Equation (2):

$$p(\mathbf{X}|\theta=1) = \prod_{n=1}^4 p(\mathbf{X}_n) \quad (5)$$

Here, each \mathbf{X}_n ($n=1, \dots, 4$) is a submatrix of \mathbf{X} defined by Equation (2), and n_k and L are similarly computed within each \mathbf{X}_n . Therefore, when $p(\theta=0) = p(\theta=1) = 1/2$, we can compute a posterior probability of θ as follows:

$$p(\theta|\mathbf{X}) \propto p(\mathbf{X}|\theta)p(\theta) \propto \begin{cases} p(\mathbf{X}|\theta=0) \\ p(\mathbf{X}|\theta=1) \end{cases} \quad (6)$$

where $p(\mathbf{X}|\theta=0)$ and $p(\mathbf{X}|\theta=1)$ are given by Equations (4) and (5), respectively. Intuitively speaking, this probability measures whether the observed score matrix is homogeneous or not.

4 Experiments

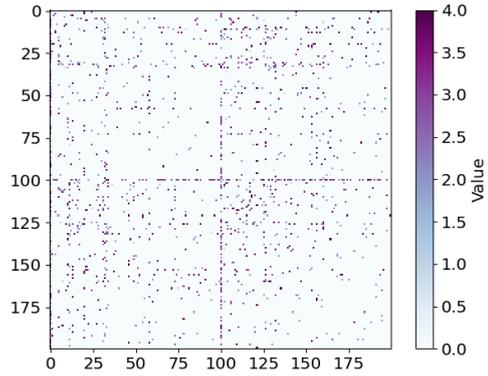
Datasets We evaluate our method on two benchmark resources: SemEval-2020 Task 1 and a subset of the WUGS datasets.²

Evaluation We evaluate all methods using Accuracy, following the standard classification protocol adopted in the WUGS framework. Each target word is labeled as stable or changed, and predictions are compared to gold-standard WUGS-derived binary labels.

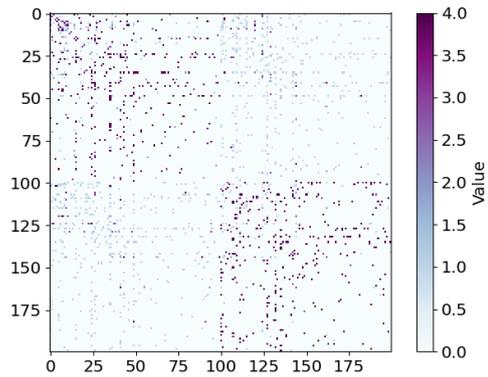
Baselines For SemEval-2020 Task 1, we compare our proposal against the three best-performing systems for each language reported in the shared task (Rother et al., 2020; Pražák et al., 2020; Asgari et al., 2020). All of these baselines rely on static or contextualized word embeddings. For WUGS datasets, where system outputs are not directly comparable across languages or domains, we adopt a simple baseline: MostFreq, which predicts the majority class (stable or changed) for each dataset.

Proposed Method To provide intuition for our method in practice, we illustrate how the proposed

²Datasets are available at <https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/wugs/>. We select only datasets with usage-usage similarity annotations, since our method requires similarity-score distributions.



(a) “Attack”



(b) “Plane”

Figure 2: Visualization of score matrices for *attack* and *plane*. *Attack* represents a word whose meaning has not changed, whereas *plane* represents a word that has undergone semantic change. In each graph, instances indexed from 0–99 correspond to the earlier period, and those from 100–199 correspond to the later period. Annotations are assigned on a scale from 1 (Unrelated) to 4 (Identical), while 0 indicates no annotation or an unknown label (Schlechtweg et al., 2021).

decision rule behaves on individual targets. Figure 2 shows the score matrices for *attack* (a semantically stable word) and *plane* (a semantically changed word).³ This figure highlights a key intuition behind our method: for semantically stable/changed words, the score matrix tends to be homogeneous/heterogeneous. The prediction is made by comparing the log-likelihoods under the **Stable** and **Changed** hypotheses (we predict **Changed** when $p(\mathbf{X}|\theta=1) > p(\mathbf{X}|\theta=0)$). For *attack*, we obtain $\log p(\mathbf{X}|\theta=0) = -6213.9$ and $\log p(\mathbf{X}|\theta=1) = -6282.8$, so the model favors the **Stable** hypothesis; the gold label is **Stable**. For

³The cross-shaped pattern in Figure 2a reflects the WUGS annotation procedure, which prioritizes informative usage pairs rather than annotating all pairs exhaustively (Schlechtweg et al., 2021). As a result, some usages are compared with many others, while many entries remain unannotated (0), producing the observed cross-shaped structure.

Data	Language	Grouping 1	Grouping 2	Accuracy (%)	
				MostFreq	Pólya
DWUG EN (Schlechtweg et al., 2021)	EN	1810–1860	1960–2010	54.3	76.1
DWUG EN Resampled (Schlechtweg et al., 2024a)	EN	1810–1860	1960–2010	60.0	80.0
DWUG DE (Schlechtweg et al., 2021)	DE	1800–1899	1946–1990	60.0	80.0
DWUG DE Resampled (Schlechtweg et al., 2024a)	DE	1800–1899	1946–1990	60.0	73.3
DiscoWUG (Kurtyigit et al., 2021)	DE	1800–1899	1946–1990	51.0	72.0
RefWUG (Schlechtweg, 2023)	DE	1750–1800	1850–1900	54.5	45.5
DURel (Schlechtweg et al., 2018)	DE	1750–1800	1850–1900	63.6	63.6
SURel (Hätty et al., 2019)	DE	general	domain specific	63.6	68.2
RuSemShift 1 (Rodina and Kutuzov, 2020)	RU	1682–1916	1918–1990	77.5	77.5
RuSemShift 2 (Rodina and Kutuzov, 2020)	RU	1918–1990	1991–2016	62.3	62.3
RuShiftEval 1 (Kutuzov and Pivovarova, 2021)	RU	1700–1916	1918–1990	74.8	74.8
RuShiftEval 2 (Kutuzov and Pivovarova, 2021)	RU	1918–1990	1992–2016	70.3	70.3
RuShiftEval 3 (Kutuzov and Pivovarova, 2021)	RU	1700–1916	1992–2016	68.5	68.5
DWUG ES (Zamora-Reina et al., 2022)	ES	1810–1906	1994–2020	55.5	78.0
DiaWUG (Baldissin et al., 2022)	ES	Spanish variant 1	Spanish variant 2	65.6	81.3
DWUG SV (Schlechtweg et al., 2021)	SV	1790–1830	1895–1903	68.1	88.6
DWUG SV Resampled (Schlechtweg et al., 2024a)	SV	1790–1830	1895–1903	60.0	73.3
ChiWUG (Chen et al., 2023)	ZH	1954–1978	1979–2003	57.5	52.5
DWUG IT (Cassotti et al., 2024)	IT	1948–1970	1990–2014	69.2	N/A
DWUG LA (Schlechtweg et al., 2021)	LA	–200–0	0–2000	55.5	N/A
NorDiaChange 1 (Kutuzov et al., 2022)	NO	1929–1965	1970–2013	67.5	75.0
NorDiaChange 2 (Kutuzov et al., 2022)	NO	1980–1990	2012–2019	77.5	70.0

Table 2: Results on WUGS datasets. For each dataset, we report accuracy for the MostFreq baseline and our Pólya-based method. The proposed approach yields strong performance across languages and dataset types, demonstrating robustness in both diachronic and synchronic SCD.

plane, we obtain $\log p(\mathbf{X}|\theta = 0) = -7427.2$ and $\log p(\mathbf{X}|\theta = 1) = -7095.6$, leading the model to favor the **Changed** hypothesis; the gold label is **Changed**. These examples illustrate the qualitative behavior of our decision rule and provide intuition for the quantitative results presented next.

Results Table 1 summarizes the results on SemEval-2020 Task 1. Despite not relying on any word embeddings, our method achieves higher accuracy than all embedding-based state-of-the-art systems in three out of four languages, demonstrating the effectiveness of modeling temporal consistency in usage similarity distributions. To assess the robustness of our framework beyond the SemEval setting, Table 2 presents results on the WUGS datasets. Our method consistently attains strong accuracy across all languages and performs robustly in both diachronic and synchronic settings. These findings indicate that the proposed framework generalizes well beyond the SemEval datasets and is applicable to diverse languages and domains.

Discussion and Future Work Although our experiments rely on human usage-similarity annotations, recent studies suggest that high-quality labels can be obtained easily through automatic methods. The DURel Annotation Tool (Schlechtweg et al., 2024b) already integrates XL-LEXEME (Cassotti et al., 2023) to automatically propose usage-

similarity annotations. Moreover, Periti and Tahmasebi (2024) show that contextualized embedding models and large language models can generate annotation labels that approach the quality of human annotations. These developments indicate that our statistical framework may be deployed in a fully automatic setting in the future, where usage-similarity labels are predicted rather than manually collected.

5 Conclusion

We proposed a statistical framework for lexical semantic change detection that models temporal consistency in usage-similarity score distributions. By evaluating whether the observed similarity structure is better explained by a single sense distribution or by distinct distributions across periods, our method provides a simple yet effective decision rule. Experiments on SemEval-2020 Task 1 and the WUGS datasets show that our approach achieves competitive or superior accuracy across languages and settings. These findings demonstrate that modeling usage-level similarity scores enables effective detection of semantic change. Together with recent advances in automatically predicting high-quality similarity labels, our framework offers a promising path toward fully automatic and interpretable semantic change detection.

Limitations

The proposed framework has several limitations that should be considered when interpreting the results and assessing its applicability.

First, our method assumes the availability of usage-level similarity scores and focuses on statistically modeling their distributions, rather than addressing how such scores are obtained. In this study, we rely on human-annotated usage-similarity labels provided by existing benchmarks. Although recent work suggests that automatic methods can generate high-quality similarity annotations (Periti and Tahmasebi, 2024), the performance of our framework under fully automatic annotation pipelines has not been empirically evaluated. In particular, automatically predicted similarity scores may exhibit different noise characteristics or systematic biases compared to human judgments, which could affect the reliability of the statistical test.

Second, related to this issue, our model treats observed similarity scores as independent samples from underlying distributions. This simplifying assumption enables an interpretable and tractable formulation, but its validity depends on the amount and structure of available data. When the number of annotations is small or unevenly distributed across usage pairs, deviations from this assumption may have a stronger impact on the estimated likelihoods. Understanding the interaction between annotation density and model reliability remains an open direction for future investigation.

Finally, our empirical evaluation is restricted to lexical semantic change detection benchmarks based on usage-similarity annotations. The applicability of the framework to other tasks or domains that can be modeled in terms of pairwise similarity judgments has not been empirically investigated. Exploring these alternative application settings would be necessary to assess the broader usefulness and generality of the proposed statistical formulation.

Acknowledgments

This work was partly supported by JST, PRESTO Grant Number JPMJPR2366, Japan, and by the NINJAL collaborative research project ‘Extending the Diachronic Corpus through an Open Co-construction Environment’.

References

- Taichi Aida and Danushka Bollegala. 2024. [A semantic distance metric learning approach for lexical semantic change detection](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7570–7584, Bangkok, Thailand. Association for Computational Linguistics.
- Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2021. [A comprehensive analysis of PMI-based models for measuring semantic differences](#). In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 21–31, Shanghai, China. Association for Computational Linguistics.
- Ehsaneddin Asgari, Christoph Ringlstetter, and Hinrich Schütze. 2020. [EmbLexChange at SemEval-2020 task 1: Unsupervised embedding-based detection of lexical semantic changes](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 201–207, Barcelona (online). International Committee for Computational Linguistics.
- Gioia Baldissin, Dominik Schlechtweg, and Sabine Schulte im Walde. 2022. [DiaWUG: A dataset for diatopic lexical semantic variation in Spanish](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2601–2609, Marseille, France. European Language Resources Association.
- Pierluigi Cassotti, Pierpaolo Basile, and Nina Tahmasebi. 2024. [DWUGs-IT: Extending and standardizing lexical semantic change detection for Italian](#). In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 190–197, Pisa, Italy. CEUR Workshop Proceedings.
- Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. [XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic change](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.
- Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. [ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection](#). In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.
- Paul Cook and Suzanne Stevenson. 2010. [Automatically identifying changes in the semantic orientation of words](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. [Analysing lexical semantic change](#)

- with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
- Anna HäTTY, Dominik Schlechtweg, and Sabine Schulte im Walde. 2019. **SURel: A gold standard for incorporating meaning shifts into term extraction**. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Minneapolis, Minnesota. Association for Computational Linguistics.
- Renfen Hu, Shen Li, and Shichen Liang. 2019. **Diachronic sense modeling with deep contextualized word embeddings: An ecological view**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. **Temporal analysis of language through neural language models**. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *WWW 2015*, pages 625–635.
- Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. **Lexical semantic change discovery**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998, Online. Association for Computational Linguistics.
- Andrey Kutuzov, Lilja Ovrelid, Terrence Szymanski, and Erik Velldal. 2018. **Diachronic word embeddings and semantic shifts: a survey**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Andrey Kutuzov and Lidia Pivovarova. 2021. **Three-part diachronic semantic change dataset for Russian**. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.
- Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. **NorDiaChange: Diachronic semantic change dataset for Norwegian**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.
- Thomas P. Minka. 2000. Estimating a Dirichlet distribution. <https://tminka.github.io/papers/dirichlet/>.
- Kevin P. Murphy. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.
- Francesco Periti and Nina Tahmasebi. 2024. **A systematic comparison of contextualized word embeddings for lexical semantic change**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.
- Ondřej Pražák, Pavel Přibán, Stephen Taylor, and Jakub Sido. 2020. **UWB at SemEval-2020 task 1: Lexical semantic change detection**. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 246–254, Barcelona (online). International Committee for Computational Linguistics.
- Julia Rodina and Andrey Kutuzov. 2020. **RuSemShift: a dataset of historical lexical semantic change in Russian**. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. **Time masking for temporal language models**. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22*, pages 833–841, New York, NY, USA. Association for Computing Machinery.
- Guy D. Rosin and Kira Radinsky. 2022. **Temporal attention for language models**. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.
- David Rother, Thomas Haider, and Steffen Eger. 2020. **CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts**. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 187–193, Barcelona (online). International Committee for Computational Linguistics.
- Dominik Schlechtweg. 2023. *Human and Computational Measurement of Lexical Semantic Change*. Ph.D. thesis, University of Stuttgart, Stuttgart, Germany.
- Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte Im Walde, and Nina Tahmasebi. 2024a. **More DWUGs: Extending and evaluating word usage graph datasets in multiple languages**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14379–14393, Miami, Florida, USA. Association for Computational Linguistics.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi.

2020. [SemEval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.
- Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. [Diachronic usage relatedness \(DUREl\): A framework for the annotation of lexical semantic change](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.
- Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. [DWUG: A large resource of diachronic word usage graphs in four languages](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldbberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte Im Walde. 2024b. [The DUREl annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta. Association for Computational Linguistics.
- Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. 2022. [Improving temporal generalization of pre-trained language models with lexical semantic change](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6380–6393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Prior and current work on semantic change*, page 51–104. Cambridge Studies in Linguistics. Cambridge University Press.
- Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. [Dynamic word embeddings for evolving semantic discovery](#). In *WSDM 2018*, page 673–681.
- Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. [LSCDiscovery: A shared task on semantic change discovery and detection in Spanish](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

The Tonogenesis Continuum in Tibetan: A Computational Investigation

Siyu Liang and Zhaxi Zerong
University of Washington
liangsy, tashi0@uw.edu

Abstract

Tonogenesis—the historical process by which segmental contrasts evolve into lexical tone—has traditionally been studied through comparative reconstruction and acoustic phonetics. We introduce a computational approach that quantifies the functional role of pitch at different stages of this sound change by measuring how pitch manipulation affects automatic speech recognition (ASR) performance. Through analysis on the sensitivity to pitch-flattening from a set of closely related Tibetan languages, we find evidence of a tonogenesis continuum: atonal Amdo dialects tolerate pitch removal the most, while fully tonal Ü-Tsang varieties show severe degradation, and intermediate Kham dialects fall measurably between these extremes. These gradient effects demonstrate how ASR models implicitly learn the shifting functional load of pitch as languages transition from consonant-based to tone-based lexical contrasts. Our findings show that computational methods can capture fine-grained stages of sound change and suggest that traditional functional load metrics, based solely on minimal pairs, may overestimate pitch dependence in transitional systems where segmental and suprasegmental cues remain phonetically intertwined.

1 Introduction

Tonogenesis refers to the emergence of lexical tone from earlier consonantal or laryngeal contrasts such as voicing and aspiration (Hombert, 1977; Haudricourt, 1954; Sæbø et al., 2025). Over time, secondary pitch perturbations can become the primary cue for distinguishing words, turning a formerly atonal language into a tonal one. For instance, in Lhasa Tibetan, the Classical Tibetan syllables *ka* (command) and *ga* (joy) were historically distinguished by aspiration and voicing; today, both are pronounced with an aspirated onset [k^h], but a high vs. low tone preserves the lexical contrast. This

process unfolds along a continuum: within a language family or group, some varieties maintain complex onset clusters and show little reliance on pitch (e.g. varieties of Amdo Tibetan), whereas others develop robust tone contrasts (e.g., Lhasa Tibetan) (Sun, 2015). Fully tonal languages (at later stages of tonogenesis) exhibit a strong dependence on fundamental frequency (f_0) for lexical contrasts, which can pose significant challenges for automatic speech recognition (ASR) if pitch information is removed (Zhang and Kirby, 2020; Fu et al., 1998).

Recent work has established that pitch-flattening—systematically removing f_0 contours while preserving spectral information—provides a computational method for quantifying tonal dependence in ASR systems (Liang and Levow, 2025). Across typologically diverse languages, fully tonal systems suffer dramatic performance degradation when pitch is removed, while non-tonal languages show minimal impact. Building on this validated methodology, we extend the investigation to multiple Tibetan languages representing distinct stages of tonogenesis. We hypothesize that flattening f_0 contours will cause a larger performance drop for languages with established tonal contrasts than for those that rely more on consonantal cues. By quantifying the impact of pitch loss on ASR using word error rate and character error rate, we offer computational evidence that captures the acoustic-level interplay between pitch and segmental cues—information that text-based functional load measures, which rely solely on minimal pair counts, cannot access. Our findings demonstrate that dialects at different points in the tonogenesis continuum exhibit correspondingly distinct degrees of pitch dependence, providing new empirical support for understanding how languages transition from consonant-based to tone-based lexical systems.

2 Literature Review

2.1 Tone and tonogenesis

Tonal languages use pitch (fundamental frequency, f_0) as a primary cue for lexical contrasts—that is, to distinguish words that would otherwise be identical. For example, in Cantonese, the syllable [si] with a high-level tone means ‘poem,’ whereas the same segments with a low-falling tone mean ‘time’ (Yip, 2002; Matthews and Yip, 2013). In contrast, non-tonal languages like English employ pitch mainly for intonation and stress (Edwards and Beckman, 1988). Crucially, tonogenesis can arise when older segmental cues (e.g., voicing or coda consonants) induce pitch perturbations that eventually become the main phonemic signal for lexical differentiation (Haudricourt, 1954; Hombert et al., 1979). In Vietnamese, for instance, the loss of final stops led to newly phonologized tones (Thurgood, 2002), while other Southeast Asian languages display partial or incomplete shifts (Matisoff, 1973).

2.2 Tibetan languages

More than fifty distinct varieties have been identified as part of the Tibetic branch of the Sino-Tibetan language family (Tournadre, 2014). Among these, the three most widely discussed groups are Central, Amdo, and Khams, illustrated in Figure 1 (Gesang and Gesang, 2002). The long-standing debate over whether these varieties constitute separate languages or dialects remains relevant (Haugen, 1966). From a strictly linguistic standpoint, many varieties are only marginally mutually intelligible (Driem, 2001), suggesting that they could be treated as separate languages. Nevertheless, socio-political and cultural factors often motivate their classification as “Tibetan dialects,” highlighting a shared literary heritage (based on Classical Tibetan) and a unified Tibetan ethnic identity (Tournadre, 2014; Gesang and Gesang, 2002). In this work, we use the term *Tibetan languages*, while acknowledging the nuanced nature of the debate.

Tibetan languages exhibit a diverse range of tonogenesis outcomes. Central (Ü-Tsang) varieties such as Lhasa Tibetan have developed robust tonal distinctions from historical voicing contrasts: modern Lhasa has two phonemic tones that yield four different pitch contours depending on syllable structure (DeLancey, 2017; Lim, 2018). Khams varieties occupy an *intermediate* position, relying partially on pitch while retaining residual segmen-

tal cues such as voicing contrasts or breathiness (Sun, 2015; Suzuki, 2011). For instance, Dege Khams preserves historically voiced consonants alongside four emerging tonal categories (Gesang and Gesang, 2002). At the opposite end of the spectrum, Amdo varieties remain primarily atonal, retaining a rich inventory of consonantal clusters in lieu of pitch-based contrasts (Gesang and Gesang, 2002; Sun, 1986).

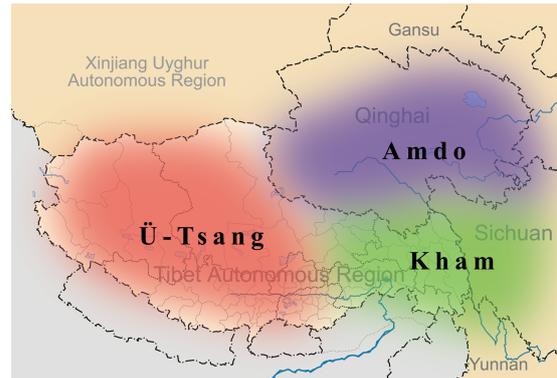


Figure 1: Approximate distribution of major Tibetan dialect groups. Ü-Tsang (red), Khams (green), and Amdo (purple) show distinct historical trajectories in tonogenesis.

2.3 Tone and ASR

Although pitch variations occur in all spoken languages for prosodic or paralinguistic reasons, *tone languages* uniquely depend on fundamental frequency (f_0) as a primary lexical contrast. As a result, failing to model pitch cues in automatic speech recognition (ASR) can lead to frequent confusions among words that differ only in tone (Fu et al., 1998; Patel et al., 2010; Wang et al., 2013; Zhang and Kirby, 2020). In practice, researchers typically address tone modeling in two broad ways. Some approaches add time-aligned pitch or voice-quality features directly into the acoustic model (Lei et al., 2006; Li et al., 2011), allowing it to learn the correlations between pitch contours and word identities. Other approaches explicitly encode tonal information in the transcript by assigning diacritics or numeric labels to otherwise identical syllables or phonemes (Yuan et al., 2021; Coto-Solano, 2021). Regardless of the chosen method, extensive evidence shows that removing or flattening pitch severely degrades ASR performance in tonal languages, mirroring the disruption that human listeners experience with f_0 suppression (Fu et al., 1998; Niu et al., 2019). In other words, in fully (or

partly) tonal systems, capturing pitch movements becomes essential for correctly mapping acoustic signals to lexical items.

Recent work has demonstrated that systematically removing pitch through flattening provides a computational method for quantifying how heavily languages rely on f_0 for lexical disambiguation: fully tonal systems suffer severe ASR degradation when pitch is removed, while non-tonal languages show minimal impact (Liang and Levow, 2025). This methodology offers a way to assess the functional load of tone in ASR systems, complementing traditional linguistic measures based on minimal pair counts.

2.4 Functional Load of Tone

In phonetics and phonology, the *functional load* of a contrast refers to how crucial that contrast is for distinguishing words (Surendran and Levow, 2004). In a strongly tonal language such as Mandarin or Cantonese, this notion conceptually relates to the amount of minimal pairs that depend solely on tone or to gauging how much pitch contributes to speech intelligibility. High functional load corresponds to greater confusion when f_0 is removed or altered, which in turn leads to marked increases in ASR error rates.

However, this measure can be inadequate for tonogenesis languages, which are still shifting from consonant-based distinctions (e.g. voicing or glottal stop codas) to primarily pitch-based contrasts (Hombert et al., 1979; Ratliff, 2015). In many transitional systems, tone can remain partially predictable from segmental cues, coarticulation, or laryngeal features (Hu, 2012; Sun, 2015). For example, pitch contours might be strongly influenced by the presence of voiced versus aspirated onsets (Hombert, 1977), or by residual glottalization that masks or enhances emerging pitch distinctions (Kirby et al., 2022). These extra cues mean that a simple tally of tone-based minimal pairs overestimates how much f_0 alone is responsible for contrasts. Even after f_0 becomes phonemic, features like breathiness or preaspiration sometimes continue to overlap with pitch distinctions, complicating the idea of an independent “tone functional load.” Consequently, although functional load remains a productive concept for mature tone languages, a single pitch-based metric does not fully capture the nuanced interplay between pitch and segmental cues in incipient or partial tone systems. In these contexts, more holistic approaches are

needed to account for how tone interacts with residual voicing, glottal stops, and other acoustic features—both for phonological description and for building robust ASR models.

3 Methods

We employ a pitch-flattening methodology to investigate the tonogenesis continuum in Tibetan varieties. This method allows us to quantify how heavily each dialect relies on pitch for lexical contrasts by measuring ASR performance degradation when f_0 information is systematically removed.

3.1 Data

We examine six Tibetan varieties drawn from TIBMD@MUC (Zhao et al., 2020), described in Table 1. Recordings featured both formal and colloquial read speech, recorded at 16 kHz, transcribed in Tibetan script. In each variety, we sampled approximately two hours of audio for training. We attempted to limit training data to two speakers per variety; in Changdu and Dege, however, three speakers were used due to data constraints. From each variety, we also reserved around 30 minutes of speech to form a test set. Data from Yushu were excluded due to insufficient data.

Since Tibetan orthography relies on stacked letters and diacritics, we converted the script to Wylie transliterations via pyewts (Wylie, 1959). We then segmented each Wylie transcription at the character level (treating spaces as distinct symbols) to form a language-specific vocabulary. Audio files were resampled at 16 kHz, and we fine-tuned separate XLS-R 300m models (Babu et al., 2021), a multilingual self-supervised speech model pretrained on 128 languages, using standard hyperparameters for low-resource ASR.

3.2 Model Training and Evaluation

We fine-tuned individual XLS-R 300m models for each of the six Tibetan varieties. During pre-processing, we normalized text and generated a character-based vocabulary for CTC training. Following model convergence, we evaluated each model on both the original test set and a pitch-flattened version of the same data.

Pitch flattening was performed using Praat’s Pitch-Synchronous OverLap and Add (PSOLA) algorithm, which replaces each utterance’s natural f_0 contour with its mean pitch. This manipulation effectively removes lexical tone cues while preserving the spectral envelope and temporal structure

Group	Duration (hrs)	Spkr	Utterances
Amdo			
Xiahe	4.12	2	3549
Aba	8.16	2	6546
Qinghai	13.65	19	12859
Kham			
Changdu	2.79	7	2558
Dege	2.31	3	1245
Yushu	0.77	3	631
Ü-Tsang			
Lhasa	37.38	48	30349
Shigatse	15.15	4	10729
Total	84.33	88	68466

Table 1: Statistics of the TIBMD@MUC database, illustrating eight Tibetan languages grouped into Amdo, Kham, and Ü-Tsang.

of the speech signal. By comparing character error rate (CER) and word error rate (WER) before and after flattening, we obtain a measure of how much each variety’s ASR system relies on pitch for lexical disambiguation.

Because Ü-Tsang varieties (e.g., Lhasa) are further along the tonogenesis trajectory, we hypothesized they would suffer greater performance drops upon pitch removal, reflecting a heavier functional load for f_0 than in the more atonal Amdo dialects. Kham varieties, described linguistically as occupying an intermediate position, were expected to show moderate sensitivity to pitch manipulation.

4 Results

Table 2 presents ASR outcomes for six Tibetan languages under original vs. pitch-flattened conditions, revealing clear differences in how each variety responds to the removal of f_0 . Amdo dialects (Xiahe, Aba), generally described as *atonal* or pre-tonogenetic (Sun, 1986; Gesang and Gesang, 2002), exhibit only moderate increases in error rates after pitch flattening, suggesting that segmental contrasts (e.g., voiced vs. voiceless onsets) continue to carry the primary lexical burden. By contrast, the Ü-Tsang group (Lhasa, Shigatse), which has been characterized as fully tonal (DeLancey, 2017), incurs notably larger performance drops. Shigatse in particular shows a sharp spike in WER (+13.9%), indicating that obscuring pitch removes a principal mechanism of word-level contrast, much as in other mature tone systems (Lim, 2018; Kirby, 2014).

Falling between these two poles are the Kham varieties (Changdu, Dege), which illustrate a partially developed reliance on pitch. Changdu undergoes a moderate rise in CER and WER when pitch is flattened, implying that f_0 is emerging as a contrastive feature yet does not fully supersede residual segmental or phonation-based cues (Sun, 2015). Dege, despite having a higher baseline error rate overall, exhibits comparatively small deltas under pitch removal, suggesting that its lexical contrasts may still be anchored in non-pitch features, or that limited training data has led to underrepresentation of the language’s tonal cues. Together, these outcomes corroborate the view that Kham varieties have not completed the shift toward fully developed tone systems (Suzuki, 2011; Kirby et al., 2022).

Overall, this pattern of results supports the broader idea of a *tonogenesis continuum* in Tibetan: atonal Amdo dialects appear minimally impacted by pitch manipulation, Ü-Tsang’s well-established tones produce sizable recognition deficits, and Kham falls in between, reflecting an incomplete transition away from historical voicing contrasts toward f_0 -based distinctions. These findings underscore the importance of capturing both segmental and suprasegmental factors in Tibetan ASR. They also suggest that, as a dialect shifts deeper into tonogenesis, the stakes of losing pitch cues rise commensurately, culminating in fully tonal systems whose lexical contrasts become acutely vulnerable to f_0 flattening.

5 Discussion

Our experiments offer insight into how pitch-flattening impacts ASR across Tibetan dialects at different stages of tonogenesis. Table 2 reveals a clear gradient pattern: the atonal Amdo varieties (Xiahe, Aba) experience only small increases in CER and WER upon pitch-flattening, pointing to a reliance on residual consonantal cues (e.g. complex onset clusters, voicing) rather than f_0 (Gesang and Gesang, 2002; Sun, 1986). At the other end of the tonogenesis continuum, the fully tonal Ü-Tsang dialects (Lhasa, Shigatse) incur larger recognition penalties, indicating that pitch has become a primary mechanism for lexical contrast (DeLancey, 2017; Lim, 2018).

The Kham varieties (Changdu, Dege) occupy an intermediate position consistent with descriptions of partial tonogenesis in which pitch has begun to assume a contrastive load but still coexists with

Language	Tone Status	Original		Flattened		Δ	
		CER	WER	CER	WER	CER	WER
Amdo							
Xiahe	Non-tonal	11.4	32.0	13.9	37.8	2.5	5.8
Aba	Non-tonal	18.2	52.5	20.2	56.3	2.0	3.8
Kham							
Changdu	Intermediate	24.7	52.3	30.3	61.3	5.6	9.0
Dege	Intermediate	47.5	90.2	49.2	91.7	1.7	1.5
Ü-Tsang							
Lhasa	Tonal	17.7	48.6	23.7	59.3	6.0	10.7
Shigatse	Tonal	17.5	49.0	25.0	62.9	7.5	13.9

Table 2: Character error rate (CER) and word error rate (WER) in % for six Tibetan dialects under original vs. pitch-flattened conditions, along with their respective Δ (Flattened–Original). Dialects are ordered along the tonogenesis continuum from atonal (Amdo) through intermediate (Kham) to fully tonal (Ü-Tsang).

other cues (Sun, 2015; Suzuki, 2011). Changdu exhibits moderate performance drops (+5.6% CER, +9.0% WER), suggesting that both pitch and residual segmental features remain important. Dege shows a smaller delta despite a higher overall error rate, possibly indicating that the dialect’s emergent tonal distinctions are underpinned by breathiness or voicing contrasts, or that limited training data dampened the model’s reliance on f_0 cues. Notably, Shigatse shows the largest WER increase (+13.9%), likely reflecting its well-documented four-way tonal system where pitch carries substantial lexical load with fewer redundant segmental cues than Lhasa (Lim, 2018). Taken together, these patterns strengthen the conclusion that Tibetan dialects vary in pitch dependence according to their historical stage of tonogenesis, mirroring prior research highlighting the gradual shift from archaic voicing contrasts to full-blown pitch-based systems (Haudricourt, 1954; Hombert et al., 1979; Ratliff, 2015).

An interesting fact of Tibetan languages is the use of a shared script dating back to Classical Tibetan. This orthography encodes older consonant clusters and voicing distinctions, thus leaving tonal contrasts unmarked for modern varieties as a result of diachronic development (Wylie, 1959). Consequently, the ASR models must learn any pitch-based distinctions directly from acoustic input. For dialects like Amdo, which retain many of the historical segmental cues, pitch contributes little to lexical disambiguation and can be removed with minimal impact on recognition. In fully tonal dialects like Lhasa, however, sound change has reduced or eliminated these consonantal cues, render-

ing f_0 indispensable for word identification (Lim, 2018).

From a typological standpoint, these results shed fresh light on how functional load interacts with partial or incipient tone. The notion of *functional load* proposes that the more minimal pairs rely on a given contrast, the more catastrophic it is to remove that contrast from the signal (Surendran and Levow, 2004). However, in a transitional system such as Kham, pitch may be correlated with, or redundant to, residual voicing or breathiness cues (Sun, 2015; Brunelle and Kirby, 2016). As a result, simply counting tone-based minimal pairs can overstate how often pitch alone distinguishes lexical items (Hombert et al., 1979; Ratliff, 2015). Our work shows that flattening f_0 does degrade recognition in Kham, but not universally to the extent observed in fully tonal dialects. Thus, while functional load captures *some* of the story, it cannot fully explain how segmental and suprasegmental cues get dynamically reweighted throughout tonogenesis. A purely text-based estimate of tone’s load would miss the subtle interplay between pitch and other laryngeal features in transitional dialects (Hu, 2012; Kirby et al., 2022).

Overall, these findings highlight the practical challenge for Tibetan ASR of capturing both the legacy of archaic consonant distinctions and the emergent or fully developed reliance on pitch. As Zhao et al. (2020); Qin et al. (2022) have shown, new acoustic features such as voice quality and breathiness may improve robustness for dialects in flux. A next step is a more granular error analysis, focusing on whether certain segments or syllables are more prone to confusion when f_0 is removed.

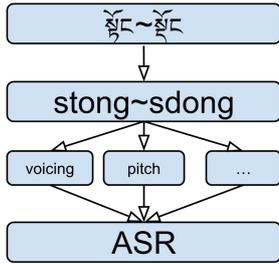


Figure 2: A schematic illustrating how Tibetan script maps to Wylie transliteration (Wylie, 1959), then to acoustic features such as pitch and voicing, which the ASR model weighs during decoding.

Such studies could clarify how partial tone systems, where voicing and pitch contrasts remain intertwined, may ultimately evolve into the strongly pitch-dependent systems exemplified by Lhasa. In addition, these insights could motivate new approaches to multi-dialect Tibetan ASR, perhaps by selectively conditioning on pitch or phonation cues only where they carry crucial lexical information.

In sum, our experiments confirm that tone systems operate along a broad continuum, from atonal to partly tonal to fully tonal. By mapping each Tibetan language’s performance profile under pitch-flattening, we obtain evidence for how the process of tonogenesis reshapes a language’s reliance on f_0 . The results not only corroborate long-standing theories of tone development (Matisoff, 1999; Hombert et al., 1979; Haudricourt, 1954), but also point to new computational challenges. For emerging tone systems, we must move beyond static estimates of functional load to account for how segmental and suprasegmental cues blend to encode lexical meaning.

6 Conclusion

We have shown that pitch-flattening reveals systematic variation in ASR performance across Tibetan varieties that directly corresponds to their position along the tonogenesis continuum. Amdo dialects (e.g., Xiahe, Aba) incurred minimal penalty from pitch suppression, reflecting their atonal nature and continued reliance on consonantal contrasts. Ü-Tsang dialects (Lhasa, Shigatse) suffered sharp error increases that underscore the functional load of pitch for lexical distinctions in fully tonal systems. Kham varieties (Changdu, Dege) fell between these extremes, showing partial reliance on f_0 alongside residual voicing or phonation cues.

From an ASR standpoint, these findings rein-

force that a one-size-fits-all acoustic modeling approach may be insufficient for languages undergoing or nearing completion of tonogenesis. Systems trained on orthographies that only encode historical clusters or voicing may inadvertently overlook emergent pitch contrasts, or conversely fail to leverage lingering segmental cues in transitional dialects. By showing that pitch removal disproportionately degrades recognition for tonal varieties, our results quantify how much current models already rely on f_0 , and suggest that this reliance varies predictably with a dialect’s stage of tonogenesis. These observations motivate future work on multi-dialect Tibetan ASR that explicitly accounts for the shifting balance between consonantal and pitch-based cues across the tonogenesis continuum.

Overall, our work highlights how pitch manipulation offers a straightforward window on the evolving functional load of tone. While techniques like measuring minimal pairs or examining segmental environments can provide a textual or phonological perspective on tone, computational experiments using pitch-flattening yield concrete evidence of how critical f_0 is for modern-day lexical discrimination. In doing so, we situate Tibetan within a cross-linguistic tonogenetic trajectory that further underscores the interplay of segmental and suprasegmental features. Future research could extend these experiments by incorporating more dialects in flux, examining detailed confusion patterns for specific segments, or exploring advanced acoustic correlates (e.g., breathiness) that may serve as additional or alternative cues in emergent tone systems.

7 Limitations

Several limitations affect the scope and generalizability of our findings. First, our study examines only six Tibetan varieties, which, while representing key points on the tonogenesis continuum, cannot capture the full diversity of Tibetan dialectology. The TIBMD@MUC corpus provides valuable data but is constrained by uneven speaker representation across varieties and relatively small test sets (approximately 30 minutes per dialect).

Second, due to the realities of linguistic fieldwork, our training and test datasets often contained recordings from the same speakers. This may inflate performance estimates and limit our ability to assess model robustness to speaker variability. Future work should validate these findings with genuinely held-out speakers to better gauge gener-

alization.

Third, the shared Tibetan orthography, while linguistically interesting, presents challenges for ASR evaluation. The script encodes historical consonant clusters rather than modern phonological contrasts, and transcription conventions may vary within and across datasets. This inconsistency particularly affects our ability to assess errors in phonologically complex categories.

References

- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhota, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. *arXiv preprint*. ArXiv:2111.09296 [cs, eess].
- Marc Brunelle and James Kirby. 2016. *Tone and Phonation in Southeast Asian Languages: Tone and Phonation in Southeast Asian Languages*. *Language and Linguistics Compass*, 10(4):191–207.
- Rolando Coto-Solano. 2021. *Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri*. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.
- Scott DeLancey. 2017. Lhasa Tibetan. In *The Sino-Tibetan Languages*.
- George van Driem. 2001. *Languages of the Himalayas*. Brill.
- Jan Edwards and Mary E. Beckman. 1988. *Articulatory Timing and the Prosodic Interpretation of Syllable Duration*. *Phonetica*, 45(2-4):156–174. Publisher: De Gruyter Mouton.
- Qian-Jie Fu, Fan-Gang Zeng, Robert V Shannon, and Sigfrid D Soli. 1998. Importance of tonal envelope cues in Chinese speech recognition. *The Journal of the Acoustical Society of America*, 104(1):505–510. Publisher: Acoustical Society of America.
- Jumian Gesang and Yangjing Gesang. 2002. Zangyu fangyan gailun. *Beijing: Minzu chubanshe*.
- André-Georges Haudricourt. 1954. De l’origine des tons en vietnamien. *Journal Asiatique*, 242:69–82.
- Einar Haugen. 1966. *Dialect, Language, Nation*. *American Anthropologist*, 68(4):922–935.
- Jean-Marie Hombert. 1977. *Development of tones from vowel height?* *Journal of Phonetics*, 5(1):9–16.
- Jean-Marie Hombert, John J. Ohala, and William G. Ewan. 1979. *Phonetic Explanations for the Development of Tones*. *Language*, 55(1):37–58. Publisher: Linguistic Society of America.
- Fang Hu. 2012. Tonogenesis in Lhasa Tibetan – Towards a gestural account. *Consonant Clusters and Structural Complexity*, 26:231.
- James Kirby, Pittayawat Pittayaporn, and Marc Brunelle. 2022. *Transphonologization of onset voicing: revisiting Northern and Eastern Kmhmu’*. *Phonetica*, 79(6):591–629.
- James P. Kirby. 2014. *Incipient tonogenesis in Phnom Penh Khmer: Acoustic and perceptual studies*. *Journal of Phonetics*, 43:69–85.
- Xin Lei, Manhung Siu, Mei-Yuh Hwang, Mari Ostendorf, and Tan Lee. 2006. *Improved tone modeling for Mandarin broadcast news speech recognition*. In *Interspeech 2006*, pages paper 1752–Tue3A2O.4–0. ISCA.
- Shang-wen Li, Yow-bang Wang, Liang-che Sun, and Lin-shan Lee. 2011. *Improved tonal language speech recognition by integrating spectro-temporal evidence and pitch information with properly chosen tonal acoustic units*. In *Interspeech 2011*, pages 2293–2296. ISCA.
- Siyu Liang and Gina-Anne Levow. 2025. *Tone in Perspective: A Computational Typological Analysis of Tone Function in ASR*. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 82–92, Vienna, Austria. Association for Computational Linguistics.
- Keh Sheng Lim. 2018. *The Tonal and Intonational Phonology of Lhasa Tibetan*. Ph.D. thesis.
- James A Matisoff. 1973. Tonogenesis in southeast Asia. *Consonant types and tone*, 1(1):71–96.
- James A Matisoff. 1999. Tibeto-Burman tonology in an areal context. In *Proceedings of the symposium “Crosslinguistic studies of tonal phenomena: Tonogenesis, Japanese Accentology, and Other Topics*, pages 3–31. Tokyo: Tokyo University of Foreign Studies, Institute for the Study of
- Stephen Matthews and Virginia Yip. 2013. *Cantonese: A Comprehensive Grammar*, 2 edition. Routledge, London.
- Yadong Niu, Fei Chen, and Jing Chen. 2019. *The effect of F0 contour on the intelligibility of Mandarin Chinese for hearing-impaired listeners*. *The Journal of the Acoustical Society of America*, 146(2):EL85–EL91.
- Aniruddh D. Patel, Yi Xu, and Bei Wang. 2010. *The role of F0 variation in the intelligibility of Mandarin sentences*. In *Speech Prosody 2010*, pages paper 890–0. ISCA.

- Siqing Qin, Longbiao Wang, Sheng Li, Jianwu Dang, and Lixin Pan. 2022. [Improving low-resource Tibetan end-to-end ASR by multilingual and multilevel unit modeling](#). *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):2.
- Martha Ratliff. 2015. [Tonoexodus, Tonogenesis, and Tone Change](#). In Patrick Honeybone and Joseph Salmons, editors, *The Oxford Handbook of Historical Phonology*, page 0. Oxford University Press.
- Jackson (Ed) Sun. 2015. [Variegated tonal developments in Tibetan](#). pages 2.1M, 35–52 pages. Artwork Size: 2.1M, 35-52 pages Medium: PDF Publisher: Pacific Linguistics Version Number: 1.0.
- Jackson T-S Sun. 1986. Aspects of the Phonology of Amdo Tibetan. (*No Title*).
- Dinoj Surendran and Gina-Anne Levow. 2004. [The functional load of tone in Mandarin is as high as that of vowels](#). In *Speech Prosody 2004*, pages 99–102. ISCA.
- Hiroyuki Suzuki. 2011. Preaspiration and tonal development in Tibetan dialects of Khams, Shar and Amdo. In *Tone, accent and intonation in eastern Eurasian languages. The 18th Meeting of the Linguistic Circle for the Study of Eastern Eurasian Languages, Aoyama Gakuin University, Tokyo*, pages 9–17.
- Lilja Maria Sæbø, Eitan Grossman, and Steven Moran. 2025. [Tonogenesis: A diachronic typology](#). *Diachronica*, 42(3-4):451–478.
- Graham Thurgood. 2002. [Vietnamese and tonogenesis: Revising the model and the analysis](#). *Diachronica*, 19(2):333–363.
- Nicolas Tournadre. 2014. The Tibetic languages and their classification. *Trans-Himalayan linguistics: Historical and descriptive linguistics of the Himalayan area*, 266(1):105–29.
- Jiuju Wang, Hua Shu, Linjun Zhang, Zhaoxing Liu, and Yang Zhang. 2013. [The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility](#). *The Journal of the Acoustical Society of America*, 134(1):EL91–EL97.
- Turrell Wylie. 1959. A standard system of Tibetan transcription. *Harvard journal of Asiatic studies*, 22:261–267. Publisher: JSTOR.
- Moira Jean Winsland Yip. 2002. *Tone*. Cambridge textbooks in linguistics. Cambridge University Press, Cambridge ; New York.
- Jiahong Yuan, Neville Ryant, Xingyu Cai, Kenneth Church, and Mark Liberman. 2021. [Automatic recognition of suprasegmentals in speech](#). *arXiv preprint*. ArXiv:2108.01122 [cs].
- Yubin Zhang and James Kirby. 2020. [The role of F0 and phonation cues in Cantonese low tone perception](#). *The Journal of the Acoustical Society of America*, 148(1):EL40–EL45.
- Yue Zhao, Xiaona Xu, Jianjian Yue, Wei Song, Xiali Li, Licheng Wu, and Qiang Ji. 2020. [An open speech resource for Tibetan multi-dialect and multitask recognition](#). *Int. J. Comput. Sci. Eng.*, 22(2-3):297–304.

A Appendix

This appendix provides additional details on our fine-tuning hyperparameters for XLS-R 300m in both experiments.

A.1 XLS-R Fine-Tuning Hyperparameters

All training runs (for both Common Voice and TIBMD@MUC data) used the same set of essential hyperparameters, with only minor adjustments for batch size depending on GPU memory:

- **Model:** facebook/wav2vec2-xls-r-300m
- **CTC loss reduction:** *mean*
- **Batch Size:** 4 or 8 per device (gradient accumulation steps adjusted to keep effective batch size at 16)
- **Learning Rate:** 3×10^{-4}
- **Optimizer:** AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$
- **Warmup Steps:** 500
- **Max Steps:** 2000
- **Vocabulary Size:** based on unique characters in the training corpus (including space or | as word delimiter).

Cross-lingual Lexical Semantic Change in Romance Languages

Ana Sabina Uban^{♣,♡} Liviu P. Dinu^{♣,♡}

Anca Dinu^{♣,♡} Simona Georgescu^{♣,♡}

University of Bucharest, [♣] Faculty of Mathematics and Computer Science,

[♣] Faculty of Foreign Languages and Literatures, [♡] HLT Research Center

{auban, ldinu}@fmi.unibuc.ro,

{anca.dinu, simona.georgescu}@l1s.unibuc.ro

Abstract

We present a comprehensive analysis of lexical semantic change in the five main Romance languages (Romanian, Italian, Spanish, French and Portuguese), based on the most exhaustive database of related words in these languages. We include both cognate words and borrowings (for the first time, to our knowledge), and compute semantic shift measures using different static and contextual embedding models, as well as three different corpora. We publish¹ the obtained lists of semantic divergences across all related word pairs, compute global trends in language-level semantic divergence, and provide insights on particular study cases of highly stable and highly divergent words for different language pairs.

1 Introduction

Approximately 27% of the lexicon of Romance languages consists of words inherited from Latin, and another 40% consists of words borrowed from Latin (Reinheimer-Rîpeanu, 2001). However, the 27% represents the fundamental core of the vocabulary (consisting of commonly used concepts - e.g. family, body parts, natural elements, everyday actions, as well as pronouns, numerals, and prepositions) and has a frequency of 80% in everyday communication (Bîrlădeanu et al., 1988). We would therefore expect that their meaning has not changed – at least not significantly – from one Romance language to another. However, a quick glance at the REW (Romanisches Etymologisches Wörterbuch), a Romance languages etymological dictionary (Meyer-Lübke, 1911), shows us that faithful preservation of the Latin meaning is reserved only for a relatively small number of words that designate concrete realities, which have not changed over time and have not been subject to changes in perception (e.g. *manus* 'hand', *oculus*

'eye', *filius* 'son', *dormire* 'sleep', *bibere* 'drink', etc.). On the other hand, many words, including those from the fields mentioned above, have undergone divergent developments (e.g. Lat. *bucca* 'cheek' > It. *bocca*, Fr. *bouche*, Es. *boca*, Pt. *boca* 'mouth' vs. Rom. *bucă* 'butt cheek'; Lat. *salire* 'to jump' > It. *salire* 'climb up' / Es. *salir* 'get out', etc.). This semantic divergence characterizes many pairs of cognates (e.g. Ro. *cugeta* 'to think' / Es. *cuidar* 'to take care (of)', Ro. *vindeca* 'to cure' / Pt. *vingar* 'to take revenge', Ro. *feri* 'to avoid' / Es. *herir* 'to hurt', etc.), whose divergent evolution has been insufficiently explained at a global level in specialized studies.

Similarly, words borrowed from Latin into Romance languages may undergo changes in meaning, albeit to a lesser extent, leading to obvious pairs of deceptive cognates (Uban et al., 2025): e.g. Es. *oficio* 'profession' / Ro. *oficiu* 'public service, office'; Ro. *transcendental* 'which is above the real world' / Es. *transcendental* 'of great importance', etc. They are still cognates, although they have entered the language through scholarly transmission.

Perhaps even more unexpectedly, there are also pairs of words borrowed from one Romance language to another that undergo a certain semantic shift, either by taking on a specialized meaning from the source language, or by subsequently changing meaning, either in the target language or in the source language (e.g. Fr. *habler* 'brag, boast', borrowed from Es. *hablar* 'speak'; Fr. *dame* 'lady' borrowed into Ro. *damă* 'prostitute'). These pairs of words are the ones that will be treated here as borrowings.

From the point of view of many linguists (Dworkin, 2006; Chauveau, 2016), semantic change seems to be the most difficult area to study because it is almost impossible to establish quantifiable parameters that can be analyzed according to scientific criteria in this field. While various formalization models have been attempted in synchronic

¹<https://nlp.unibuc.ro/resources.html#HistoricalLinguistics>

semantics, in diachronic semantics, i.e., the way in which the meaning of words evolves, attempts in this direction are concentrated in the field of computational linguistics and are still fairly recent, and mostly comprise of monolingual studies.

Therefore, we aim to automatically measure the semantic divergence between cognates in any two Romance languages from the main core (Romanian, Italian, French, Spanish, Portuguese) by representing them as contextual embeddings extracted from corpora (see section *Data*). Once we have obtained the distances between any two such representations of cognates, we can compute an aggregated global distance for any two languages, thus obtaining a corpus driven semantic divergence for the Romance language family. The interpretation of these results will show to what extent geographical distance is a source of proximity between languages, as proposed in [Bartoli \(1925\)](#), who showed that languages on the periphery of the former Roman Empire change less as a result of their isolation from the center) or, on the contrary, a source of semantic distance as a result of different interpretations of reality. We assume that a computational approach could provide a significant platform for a methodologically coherent analysis of semantic change, proposing quantifiable paradigms for historical semantics that can be explored using scientific tools, which can, of course, be improved or adapted depending on the particular task.

2 Related Work

Semantic change has become an increasingly central topic in computational historical linguistics over the past decade, driven by the availability of large multilingual corpora and significant advances in distributional and contextual semantic modeling. These research relied either on distributional similarity and static embeddings to capture cross-lingual meaning variation, or, more recently, on contextualized representations, clustering-based methods. Work on semantic change comprise both traditional diachronic approaches, usually monolingual, as well as synchronic, multi-lingual ones.

[Montariol and Allauzen \(2021\)](#) introduce a computational framework for tracking semantic divergence between translated word pairs across languages and time, showing that contextualized embeddings combined with clustering outperform static representations in capturing gradual meaning shifts.

[Uban et al. \(2021\)](#) analyze semantic change within cognate sets across English and Romance languages, highlighting systematic differences between cognate words and demonstrating that lexical properties such as frequency and polysemy correlate with semantic divergence.

Building on previous approaches, [Kawasaki et al. \(2022\)](#) explicitly draw on cross-lingual divergence measures to revisit classical statistical laws of semantic change in Romance cognates, refining the roles of frequency, polysemy, morphological complexity, and lexical age in explaining semantic stability and drift.

Most recently, [Uban et al. \(2025\)](#) extend this line of research through a large-scale computational investigation of semantic false friends across Romance languages, introducing new etymologically grounded resources and evaluation protocols that further clarify how shared origin can nevertheless result in substantial cross-linguistic semantic divergence.

This work draws from previous methodologies and data, integrating the most recent datasets and methods. Our contribution is twofold: on the one hand we perform all our experiments based on the exhaustive database of Romance related words RoBoCoP ([Dinu et al., 2023](#)), including cognates and borrowings, and, on the other, we employ different models including static and contextual embeddings for those related words, separately for cognates and borrowings, from three different multilingual and parallel large corpora, for multi-level comparison.

3 Data

3.1 Cognates and Borrowings Dataset

We perform our analyses on pairs of related words extracted from the most comprehensive database of related words in Romance languages up to date, sourced from etymological dictionaries and manually curated, RoBoCoP (ROMance BORrowing COgnate Package and Benchmark for Multilingual Cognate Identification) ([Dinu et al., 2023](#)). As a source of cognate word pairs, we use the freely available subset ProtoRom ([Dinu et al., 2024a](#)), a database of cognate tuples and etymons in the five Romance languages, with 19,222 entries (tuples with at least 2 cognates). We extract borrowings from the original RoBoCoP database, totaling 46,490 borrowing pairs across Romance languages pairs ([Dinu et al., 2024b](#)).

3.2 Word Embeddings Corpora

For our computational experiments, we rely on word embeddings as models of meaning representation. In order to compare the effect of the corpus used to train the embeddings, we experiment with three different parallel corpora to extract embeddings:

- *Wikipedia*²
- *Europarl*, a standard parallel corpus with aligned sentences including the Romance languages, based on proceedings of the European Parliament (Koehn, 2005),
- *RomCro2.0*, a recent parallel corpus including more general language sourced from literary works written in various original languages and translated in Romance languages and Croatian (Mikelenić et al., 2024).

4 Methodology

Our proposed algorithms rely on representing the related words in semantic space based on word embeddings and then measuring the semantic distances between them in the obtained multidimensional space, using two different embedding algorithms:

- contextual embeddings extracted from a BERT transformer pretrained on a multilingual sentence similarity task for optimizing sentence representations, based on a SentenceBERT architecture (Reimers and Gurevych, 2019)³, as well as the multilingual transformer xlm-roberta-base (Conneau et al., 2019) for a subset of the experiments,
- static FastText aligned embeddings (Bojanowski et al., 2016; Conneau et al., 2017), which have been previously used successfully for cognate semantic divergence measures (Uban et al., 2019; Uban and Dinu, 2020; Uban et al., 2021).

We use the publicly available pre-aligned multilingual static embedding spaces based on the Wikipedia corpus⁴, obtained by training a linear transformation using the Procrustes alignment algorithm, as published in Conneau et al. (2017).

²<https://huggingface.co/datasets/wikimedia/wikipedia>

³<https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2>

⁴https://github.com/babylonhealth/fastText_multilingual/

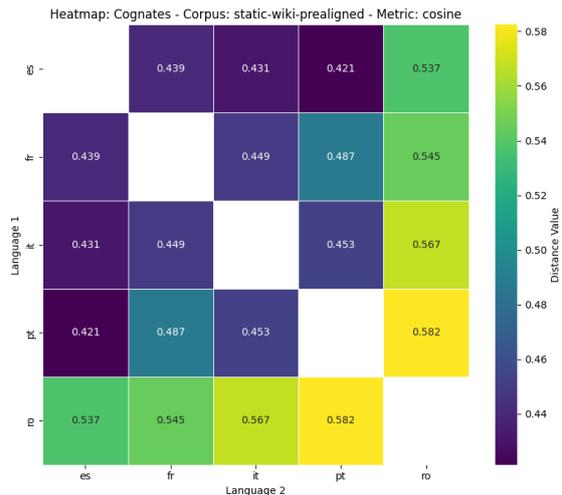


Figure 1: Semantic divergence between cognates in Romance languages based on cosine distance on static embeddings on the Wikipedia corpus

In the case of contextual representations, in order to extract unique vectorial representations for each cognate and borrowing word from the pretrained sBERT model and the three corpora, we first identify each target word in our database in the corpus, based on their stems (obtained using the Snowball stemmer). We obtain for each cognate/borrowing a set of embeddings corresponding to each occurrence in the corpus (including potentially different senses of the word), and experiment with three different methods for computing distances between cognates based on the sets of their corresponding embeddings, inspired from the best solutions proposed in (Periti and Tahmasebi, 2024):

- mean distance: a simple dimension-wise average of the embeddings is computed to obtain unique representations per cognate, then cosine similarity is used to compute distances;
- JSD: embedding clusters for each cognate are generated using affinity propagation and cosine distance, the Janson-Shannon divergence is computed between the clusters as a distance metric between cognates;
- WiDiD (Periti et al., 2022): embedding clusters for each cognate are generated independently, and cluster centers are computed using simple averaging, then the distance between clusters is computed as the cosine distance between cluster centers.

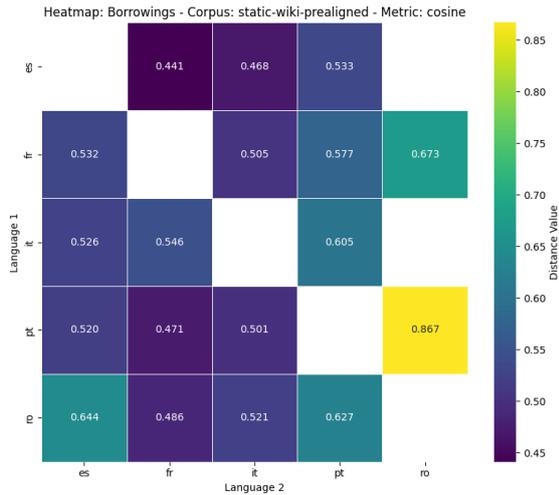
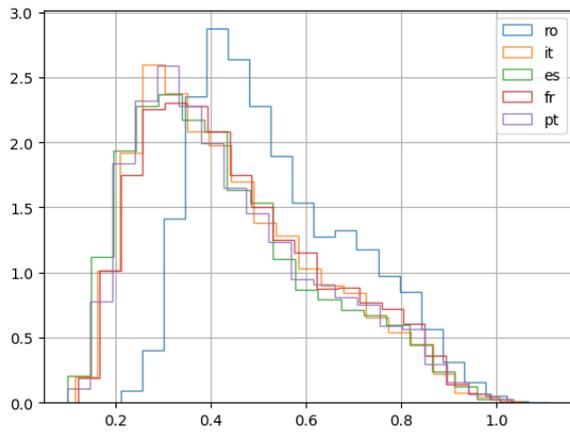
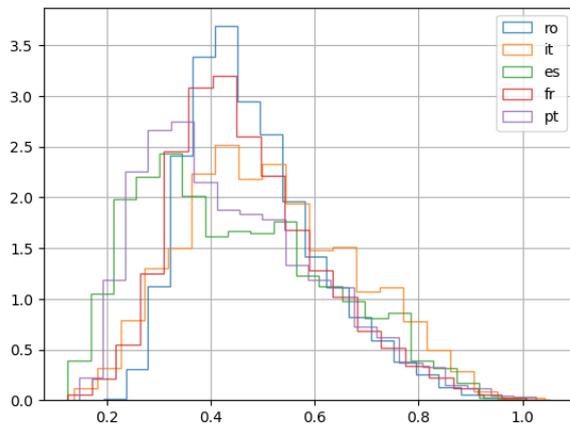


Figure 2: Semantic divergence between borrowings in Romance languages based on cosine distance on static embeddings on the Wikipedia corpus



(a) Distribution of cognates distances for each language based on Wikipedia static embeddings.



(b) Distribution of borrowings distances for each language based on Wikipedia static embeddings.

Figure 3: Distributions of cognates and borrowings distances for each language based on Wikipedia static embeddings.

5 Results

Global language divergence scores based on the Wikipedia corpus with static embeddings are shown in Figures 1 and 2 for cognates and borrowings, respectively. Figure 4 shows cognate semantic divergence computed with contextual embeddings on the three corpora. Similar heatmaps for borrowings using contextual embeddings, as well as results using additional distance metrics, are shown in the Appendix.

Figures 5 and 6 show semantic divergences based on contextual embeddings, as rankings of language pairs, from most distant to most similar. We notice few differences between the static and contextual embeddings results - the ranking of global language pair distances is generally maintained. Since the problem of evaluation is more difficult in the case of the present study, we rely on results in Uban et al. (2025) for choosing the models we choose to focus on primarily: here contextual embeddings based on Wikipedia using mean cosine distance, as well as static Wikipedia embeddings with cosine distance, are the most useful for detecting false friends.

The distribution of mean distances between related words in each language to any other language (where distances for each word in a given language to all related words, irrespective of their language, are averaged together), based on Wikipedia static embeddings, is shown in figures 3a for cognates and 3b for borrowings. We can observe that the curve for Romanian cognates is skewed to the right and more ample compared to all other Romance languages, meaning that Romanian cognates semantically diverged more than the Western cognates, most probably due to its geographical isolation in between non-Romance languages. The pattern is similar with the borrowings distribution, where Romanian borrowings changed the most, followed by French ones. Most distributions have a relatively normal shape, with a skew to the left. There is a slight multimodality in the distribution of Spanish and Portuguese borrowings, with two main peaks in the distributions around distances of 0.3 and 0.5: for both Spanish and Portuguese, the lower peak corresponds to borrowings from French (with an average distance of 0.44 for Fr-Es and 0.47 for Fr-Pt), and the higher peak to distances with each other (0.51 average distance for Es-Pt and 0.53 for Pt-Es). The individual distance distributions of borrowings from or to Spanish and Portuguese and

each of the other Romance languages are illustrated in the Appendix.

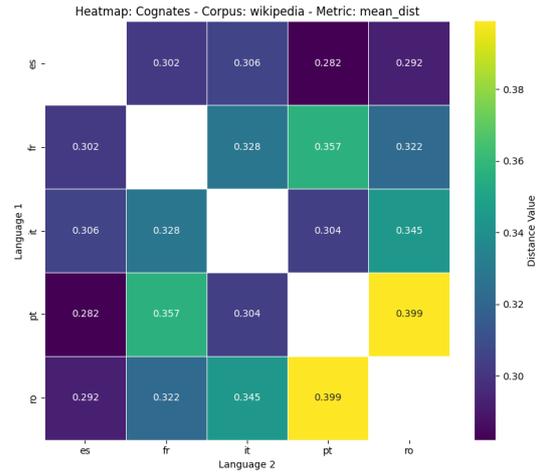
5.1 Global language divergence scores for different embedding models and corpora

For cognates, the mean distance between languages computed statically on Wikipedia and contextually on Wikipedia, Romcro, and Europarl corpora is represented in figures 5a, 5b, 5c, and 5d, respectively. The rankings of the semantic divergence between languages is dependent on the method and corpus, but the pattern is clear: the most divergent pair is Pt-Ro and the least divergent Es-Pt; on average, the pairs between Spanish and any other language show the lowest degrees of semantic differentiation, while pairs containing French are moderately divergent.

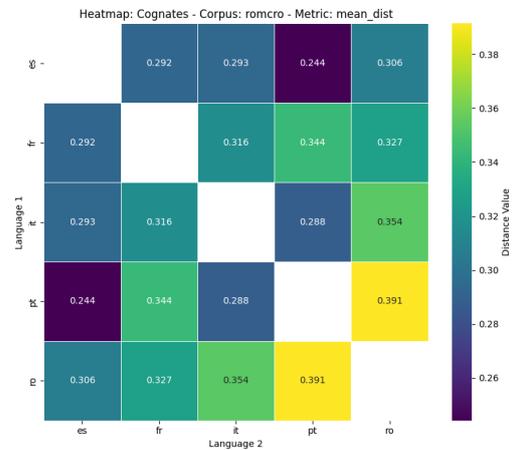
The proximity between Es-Pt in each corpus indicates not so much a greater degree of preservation in relation to Latin, but rather the fact that the two languages had a greater level of cohesion between them, a common evolution over a longer period of time and, at the same time, an evolution separate from other languages due to the geographical position of the Iberian Peninsula.

It is also noteworthy that in the static embeddings based on Wikipedia, all pairs containing Romanian show the highest level of divergence. This situation seems to reflect the effects of the isolation of the Romanian language, separated by a consistent Slavic fringe from the rest of the Romance languages, which predictably could lead to greater semantic divergence. At the same time, this distance, which appears in all corpora as the greatest, partially contradicts Bartoli’s hypothesis, according to which lateral areas share more common features with each other than with the rest of the Romance languages. However, the distance between Ro and Es is significantly smaller than in other pairs, including those from central Romania, such as Fr-It.

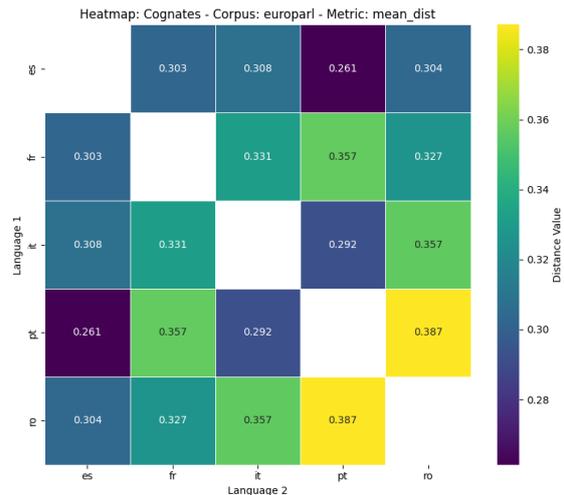
Following a comparative analysis of the corpora, it can be observed that the distances between Romanian and other languages (except for Portuguese) vary slightly, which required taking into account the specificity of the corpora. We were thus able to observe that in the corpus of parliamentary speeches, the distances are smaller because the language used is standard, specific to political and economic speeches, which leads to the use of a neological lexicon of Latin-Romance origin common to all Romance languages and, moreover, largely shared with English. In contrast, the language used



(a) Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Wikipedia (using mean-dist)



(b) Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using mean-dist)



(c) Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using mean-dist)

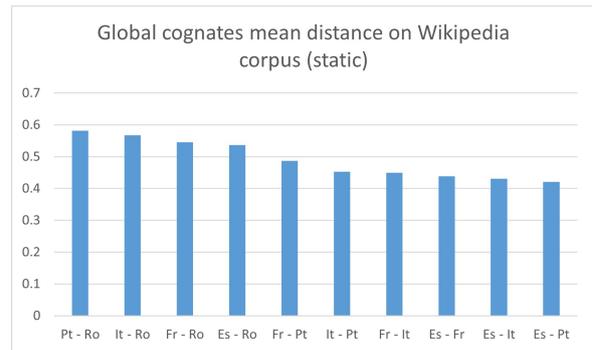
Figure 4: Heatmaps of semantic divergence scores between cognates in Romance languages based on contextual embeddings trained on three corpora (Wikipedia, RomCro, Europarl; using mean-dist).

in literature (the RomCro corpus) shows a greater variety, since the lexicon present in such texts is more diverse in terms of origin, while it does not give such a high weight to neologisms.

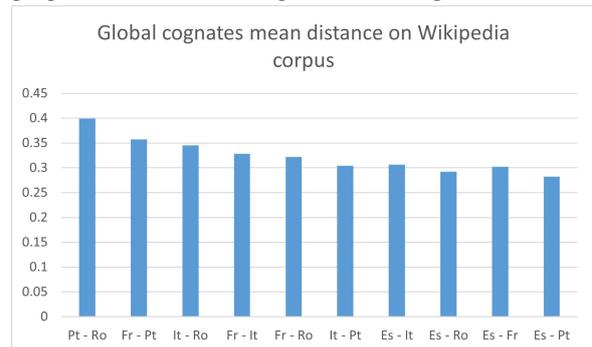
For borrowings, the number of language pairs is double, because the direction of borrowings matters. The mean semantic divergences between borrowings computed on static Wikipedia embeddings, as well as contextual embeddings on Wikipedia, Romcro, and Europarl are given in figures 6a, 6b, 6c, and 6d, respectively. When it comes to borrowing, things are much more nuanced, and patterns applicable to all languages cannot be detected, for several reasons. Firstly, the number of borrowings from one Romance language to another is incomparably smaller than cognates. Romanian will always produce unbalanced pairs because other languages have borrowed very few words from this language, which makes the data unreliable; Romanian, on the other hand, has borrowed heavily from French (approximately 9% of the Romanian vocabulary consists of French borrowings - although treated in lexicography as words of multiple etymology French/ Latin/ Italian -, a proportion that no other Romance language comes close to, cf. [Reinheimer-Rîpeanu \(2001\)](#)), but very few words from Spanish and almost none from Portuguese. Secondly, the conceptual domains from which borrowings were made are limited to certain elements that are perceived as specific to each culture, or that have penetrated a linguistic community along with the designated object (e.g. Romanian borrows from Spanish *telenovela*, *marijuana*, *cacao*; French from Italian *operetta*, *mosaico*, *miniatura*; Pt. from Es. *tabaco*, *guerrilla*, *coca*, etc.). The distances are small in the case of words that denote the same concept in the target language as they did in the source language (Fr. *chocolat*, *sangria*, It. *lama* < Es. *llama*), but large in situations where the borrowed word is taken with a single meaning selected from several that it had in the source language (Fr. *embargo* 'embargo' vs. Es. *embargo*, which is mostly used in the phrase *sin embargo* 'however', absent from Fr.).

5.2 Manual analysis of most changed and most stable related words

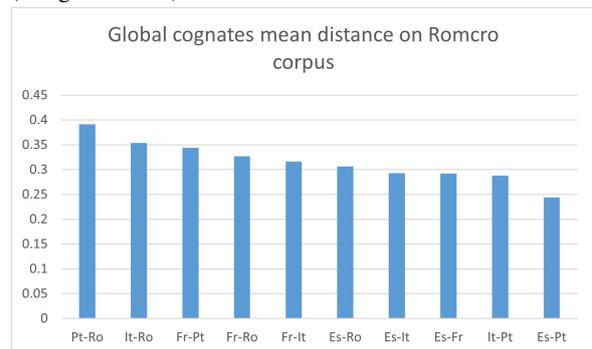
The results were manually analyzed in detail as follows. For each language pair, we sorted the cognates and borrowings according to the distance obtained and observed several patterns in terms of conceptual areas where low divergences occurred:



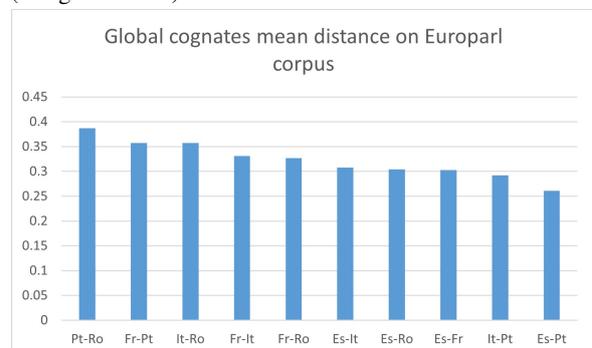
(a) Cognates semantic divergence between Romance languages based on static Wikipedia embeddings



(b) Cognates semantic divergence between Romance languages based on contextual embeddings trained on Wikipedia (using mean-dist)



(c) Cognates semantic divergence between Romance languages based on contextual embeddings trained on Romcro (using mean-dist)



(d) Cognates semantic divergence between Romance languages based on contextual embeddings trained on Europarl (using mean-dist)

Figure 5: Cognates semantic divergence between Romance languages based on different corpora (as language pair rankings).

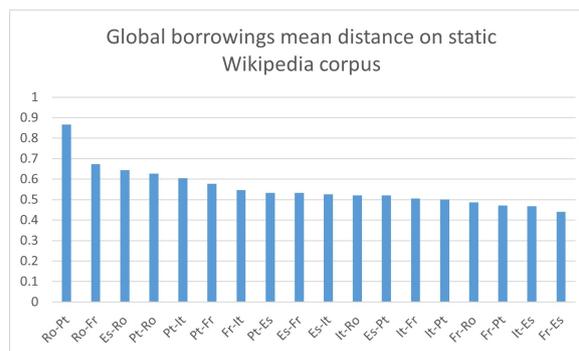
these are generally generic words that designate scientific fields or general areas of activity ('mathematics', 'astronomy', 'agriculture', 'medicine', etc.), univocal verbs, in other words, verbs that have not developed figurative meanings (to write, to kill), but also technical verbs ('to transport', 'to torture', 'to excommunicate'). At the other extreme we find terms that have either changed register (Ro. *muiere* is a regional and derogatory word for 'woman', whereas Es. *mujer* is the standard term for 'woman' and 'wife'), have restricted or expanded their area of application (e.g. Ro. *bucată* 'piece' - semantic expansion - vs Es. *bocado* 'bite' - from Lat. **buccata* 'mouthful'; Fr. *comprendre* 'to understand' - semantic narrowing - vs Ro. *cuprinde* 'to get hold of').

In most cases, the cause of the large semantic distance lies in the polysemic areas developed by cognates, which do not overlap, and therefore the terms appear in different contexts (e.g. Ro. *popor* 'people of a country' vs Es. *pueblo* 'people of a country' and 'village'). At the same time, cases of homonymy are misleading: while they can only be avoided through manual intervention, they result in the calculation of a large distance between cognates that would otherwise be semantically close: e.g. It. *aglio* 'garlic' etymologically corresponds to Ro. *ai* 'garlic' - a regional word -, which formally coincides with the the indicative 2nd pers. sg. of the verb *a avea* 'to have', incomparably more frequent: therefore, the divergence is 95% for this cognate pair; similarly, Es. *san / santo* 'saint' corresponds to Ro. *sân*, whose application is limited to contexts such as *Sân Nicolau*, *Sân Gheorghe*; otherwise, the adjective *sân* is homonymous with the noun *sân* 'breast'; in Pt-Es pair, Pt. *flama* 'flame' is the cognate of Sp. *llama* 'id.', which in its turn is homonymous with *llama* 'llama' (animal), and with the verb *llamar* 'to call', indicative 3rd pers. sg.).

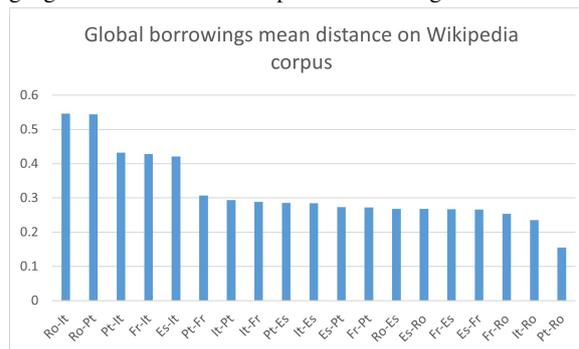
5.3 Part of speech distribution of words undergoing shifts

We separately measure semantic shifts for words with different parts of speech, according to Open multilingual WordNet. Some words can have multiple parts of speech according to WordNet - in these cases we consider them for both parts of speech. The coverage of words analyzed in WordNet is shown in the Appendix.

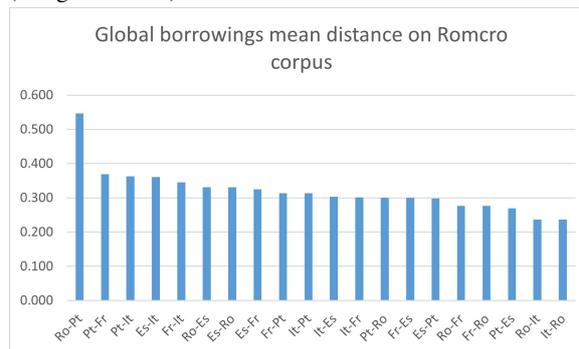
The mean cognate and borrowing distances, respectively, for each part of speech, based on



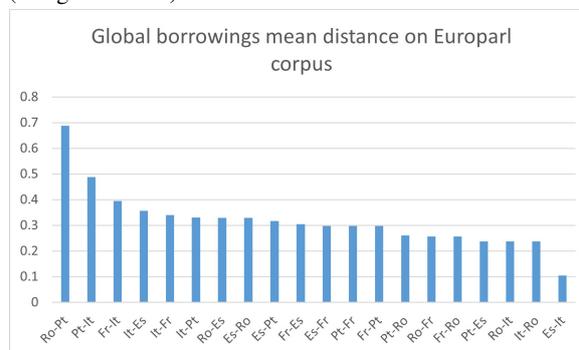
(a) Borrowings semantic divergence between Romance languages based on static Wikipedia embeddings.



(b) Borrowings semantic divergence between Romance languages based on contextual embeddings trained on Wikipedia (using mean-dist)



(c) Borrowings semantic divergence between Romance languages based on contextual embeddings trained on Romcro (using mean-dist)



(d) Borrowings semantic divergence between Romance languages based on contextual embeddings trained on Europarl (using mean-dist)

Figure 6: Borrowings semantic divergence between Romance languages across different corpora (as language pair rankings).

Wikipedia static embeddings are given in Tables 1 and 2 - shown for each language separately and overall for all words where we could extract the part of speech. The differences across POS are minor. Adjectives are most semantically stable across languages, particularly for borrowings, since borrowed adjectives are mostly relational adjectives, rather than describing a subjective quality (e.g. Fr-Es *biológico* 'biological', *informático* 'informational', Fr-Ro *casabil* 'breakable', *demonstrativ* 'demonstrative', *intestinal* 'intestinal'). For cognates, verbs are most stable overall, for most languages except for Romanian. Semantic shifts in nouns vary by language: nouns are less stable in Italian, Spanish and Portuguese.

Table 1: Mean cognates distance for each part of speech, based on Wikipedia static embeddings.

Language	POS	#words	Avg. dist.
ro	noun	6200	0.5396
ro	verb	2379	0.5480
ro	adjective	777	0.5260
ro	adverb	1076	0.5275
it	noun	7352	0.4545
it	verb	3267	0.4148
it	adjective	1256	0.4343
it	adverb	137	0.5085
es	noun	5729	0.4432
es	verb	2736	0.3970
es	adjective	1100	0.4462
es	adverb	101	0.4781
fr	noun	5892	0.4616
fr	verb	3231	0.4378
fr	adjective	1002	0.4571
fr	adverb	282	0.5197
pt	noun	6701	0.4511
pt	verb	3464	0.4170
pt	adjective	937	0.4161
pt	adverb	163	0.5398
overall	noun	31874	0.470
overall	verb	15077	0.438
overall	adjective	5072	0.452
overall	adverb	1759	0.523

6 Conclusions and Future Work

We have presented a complete analysis of lexical semantic divergence in Romance languages based on different word embeddings models trained on different corpora, including the most exhaustive vocabulary of cognates as well as borrowings in Romance languages. We find the highest semantic proximity for related words in Spanish and Portuguese, both in the case of borrowings and cognate words. Romanian generally stands out with words

Table 2: Mean borrowings distance for each part of speech, based on Wikipedia static embeddings.

Language	POS	#words	Avg. dist.
ro	noun	8068	0.4844
ro	verb	1308	0.5093
ro	adjective	797	0.4570
ro	adverb	1136	0.4632
it	noun	2040	0.5109
it	verb	298	0.5045
it	adjective	296	0.5218
it	adverb	35	0.7112
es	noun	775	0.4527
es	verb	98	0.3975
es	adjective	52	0.4643
es	adverb	1	0.5784
fr	noun	9755	0.4737
fr	verb	2251	0.4890
fr	adjective	1737	0.4368
fr	adverb	103	0.5347
pt	noun	1340	0.4525
pt	verb	113	0.4022
pt	adjective	88	0.3714
pt	adverb	5	0.5318
overall	noun	21978	0.470
overall	verb	4068	0.492
overall	adjective	2970	0.449
overall	adverb	1280	0.476

diverging most from their cognates in other romance languages, with Romanian and Portuguese being the most distant language pair overall, in contradiction to Bartoli’s lateral areas hypothesis. We find some differences in the rankings of language pairs based on average related word divergence, due to differences in vocabulary as well as divergence tendencies across language registers, confirming that including additional spoken corpora might be a useful complement to our results.

In the future, refining the contextual embedding representations by post-alignment of embedding spaces across languages could improve the precision of the resulted distance measures. Handling polysemy and treating each word sense separately could offer additional interesting insights.

Limitations

While we rely on three different multilingual corpora in different domains to obtain a complete perspective on the usage of the words analysed in different languages and contexts, complementing these with spoken language or social media corpora might be useful for capturing more subtle colloquial or metaphorical meanings.

The manual analysis of word pairs with high

semantic distance has revealed a small number of dictionary errors which might introduce some noise in the results.

Ethical Statement

There are no ethical issues that could result from the publication of our work. Our experiments comply with all license agreements of the data sources used. We make the contents of our package available for research purposes.

Acknowledgements

This research was supported by the Ministry of Education and Research, CNCS-UEFISCDI, project SIROLA, number PN-IV-P1- PCE-2023-1701, within PNCDI IV, and by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization, and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

References

- Matteo G. Bartoli. 1925. *Introduzione alla neolinguis-tica: Principi, scopi, metodi*. Olschki, Firenze.
- Mihaela Bîrlădeanu, M. Iliescu, Liliana Macarie, Ioana Nichita, Mariana Ploae-Hănganu, Marius Sala, Maria Theban, and Ioana Vintilă-Rădulescu. 1988. *Vocabularul reprezentativ al limbilor romanice*. Editura Științifică și Enciclopedică, București.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.
- Jean-Paul Chauveau. 2016. Reconstruction comparative et histoire sémantique. In Éva Buchi and Wolfgang Schweickard, editors, *Dictionnaire Étymologique Roman (DÉRom) 2. Pratique lexicographique et réflexions théoriques*, pages 53–65. De Gruyter, Berlin/Boston.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#). *CoRR*, abs/1911.02116.
- Alexis Conneau, Guillaume Lample, Marc Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.
- Liviu P Dinu, Ana Uban, Alina Cristea, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023. Robocop: A comprehensive romance borrowing cognate package and benchmark for multilingual cognate identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7610–7629.
- Liviu P Dinu, Ana Uban, Alina Cristea, Ioan-Bogdan Iordache, Teodor-George Marchitan, Simona Georgescu, and Laurentiu Zoicas. 2024a. Verba volant, scripta volant? don't worry! there are computational solutions for protoword reconstruction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6314–6326.
- Liviu P Dinu, Ana Uban, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2024b. It takes two to borrow: a donor and a recipient. who's who? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6023–6035.
- Steven N. Dworkin. 2006. Recent developments in spanish (and romance) historical semantics. In *Selected Proceedings of the 8th Hispanic Linguistics Symposium*, pages 50–57, Somerville. Cascadilla Proceedings Project.
- Yoshifumi Kawasaki, Maëlys Salingre, Marzena Karpinska, Hiroya Takamura, and Ryo Nagata. 2022. [Revisiting statistical laws of semantic shift in romance cognates](#). In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, pages 141–151, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.
- Wilhelm Meyer-Lübke. 1911. *Romanisches etymologisches wörterbuch*, volume 3. C. Winter.
- Bojana Mikelenić, Antoni Oliver, and Marko Tadić. 2024. Expansion of the romcro corpus with texts in catalan. In *CLARIN Annual Conference Proceedings 2024*, pages 135–139. Barcelona: CLARIN.
- Syrielle Montariol and Alexandre Allauzen. 2021. [Measure and evaluation of semantic divergence across two languages](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1247–1258, Online. Association for Computational Linguistics.
- Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. [What is done is done: an incremental approach to semantic shift detection](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 33–43, Dublin, Ireland. Association for Computational Linguistics.
- Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. *CoRR*.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sanda Reinheimer-Rîpeanu. 2001. *Lingvistica romanică. Lexic – morfologie – fonetică*. All, Bucarest.

Ana Sabina Uban, Alina Maria Ciobanu, and Liviu P Dinu. 2019. Studying Laws of Semantic Divergence across Languages using Cognate Sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166.

Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Liviu P. Dinu, Simona Georgescu, and Laurentiu Zoicas. 2021. [Tracking semantic change in cognate sets for english and romance languages](#). In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change*, pages 64–74, Online. Association for Computational Linguistics.

Ana Sabina Uban and Liviu P Dinu. 2020. Automatically Building a Multilingual Lexicon of False Friends With No Supervision. In *Proceedings of LREC 2020*, pages 3001–3007.

Ana Sabina Uban, Liviu P. Dinu, Bogdan Iordache, Simona Georgescu, and Claudia Vlad. 2025. [Friend or foe? a computational investigation of semantic false friends across romance languages](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15310–15324, Online. Association for Computational Linguistics.

A Appendix

A.1 Infrastructure and libraries

The experiments were performed on an RTX 2080 Ti GPU and a Ryzen 5 3600X CPU. Libraries used for embedding extraction, cognate and corpora pre-processing (extracting stems), synonym extraction based on WordNet, and distance metrics computation:

- keras==3.8.0
- keras-hub==0.18.1
- keras-nlp==0.18.1
- nltk==3.9.1
- scikit-learn==1.6.1
- scipy==1.13.1
- sentence-transformers==3.4.1
- spacy==3.7.5

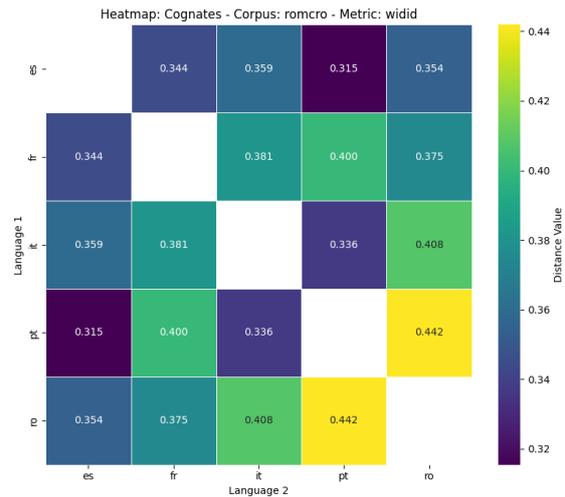


Figure 7: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using Widiid)

- tensorflow==2.18.0
- tensorflow-datasets==4.9.7
- transformers==4.48.3
- and fasttext vector support based on https://github.com/babylonhealth/fastText_multilingual/.

Transformer models used:

- distiluse-base-multilingual-cased-v2: 135M parameters
- xlm-roberta-base: 279M parameters

Hyperparameters:

- maximum number of sampled occurrences for a word when computing contextual embeddings: 200
- occurrence matching was checked based on stem matching with and without unicode normalization (removing of accents)
- Affinity Propagation clustering was trained with the default hyperparameters provided by the scikit-learn library.

A.2 Additional Results

- IT-ES: total: 3666, not in WordNet (WN): 1923
- IT-FR: total: 2172, not in WN: 918

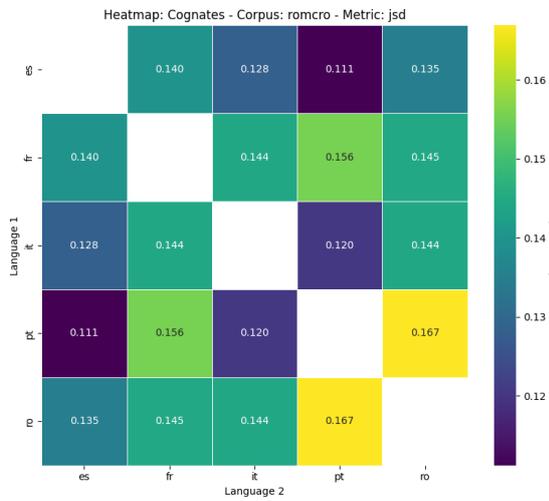


Figure 8: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using JSD)

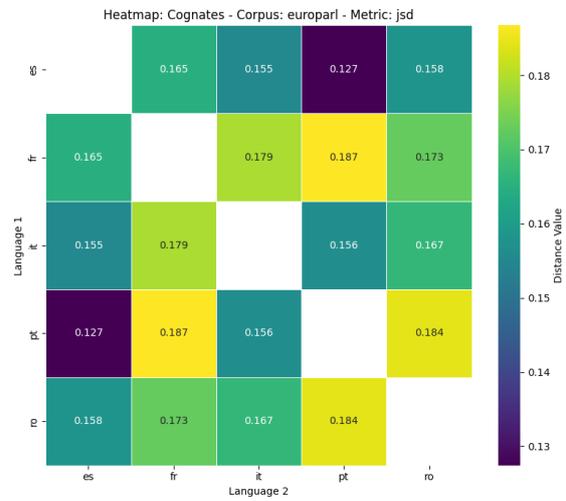


Figure 10: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using JSD)

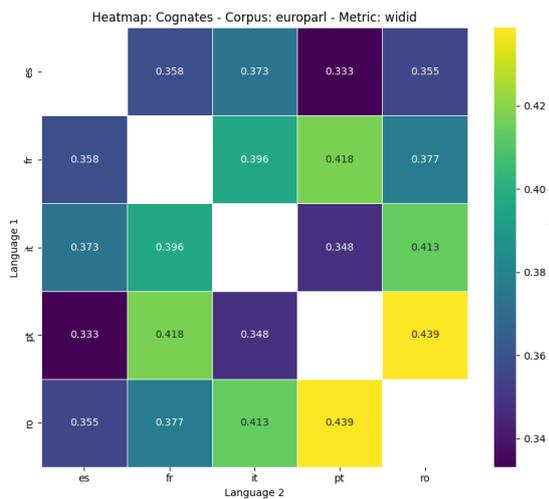


Figure 9: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using widid)

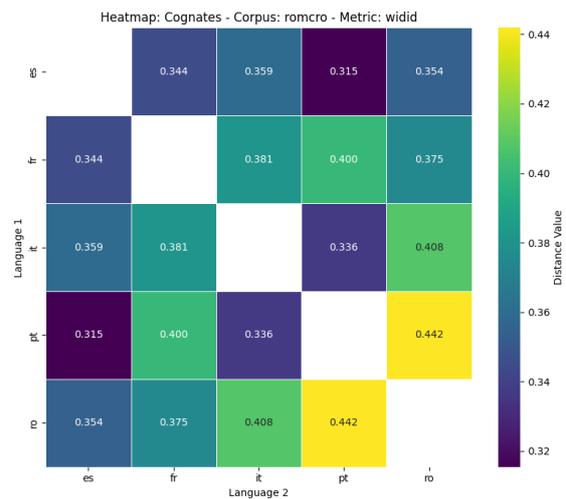


Figure 11: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using Widid)

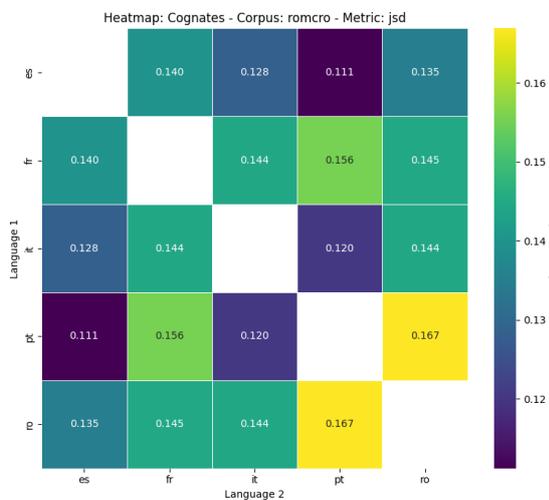


Figure 12: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using JSD)

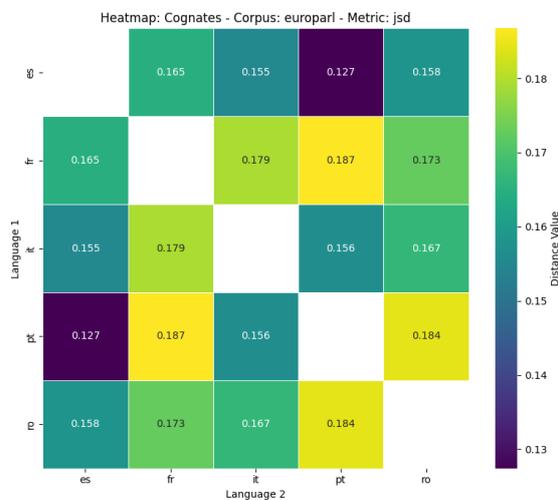


Figure 14: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using JSD)

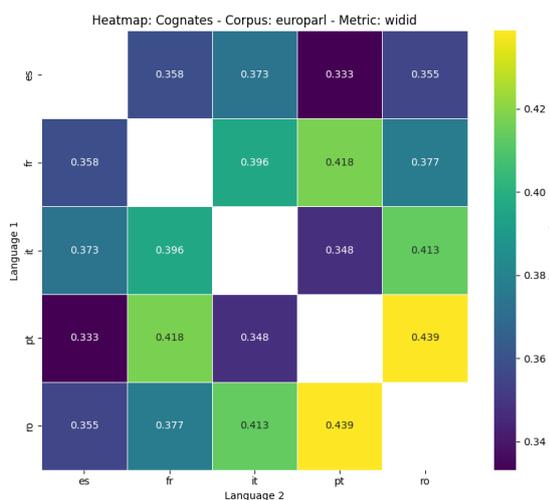


Figure 13: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using widid)

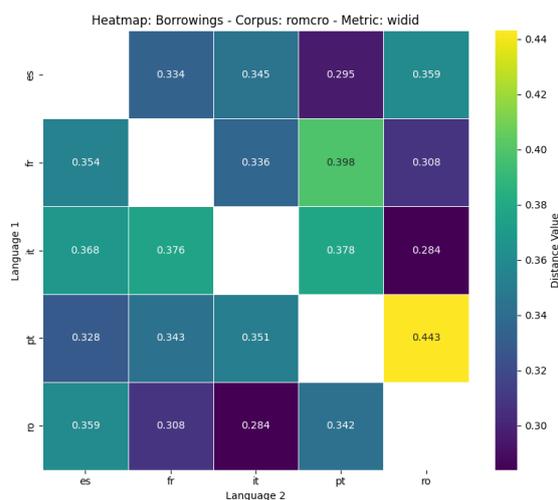


Figure 15: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on RomCro (using Widid)

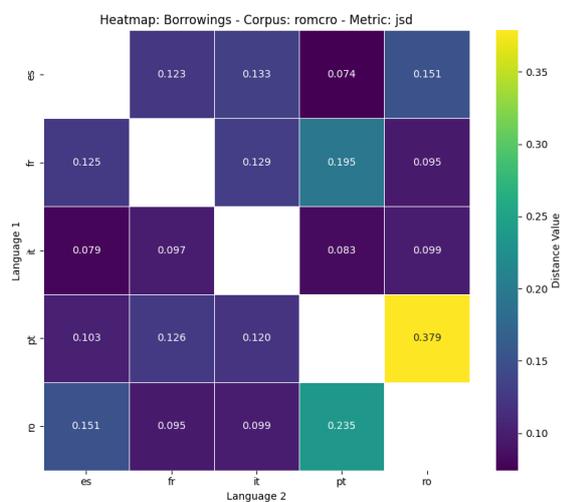


Figure 16: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on RomCro (using JSD)

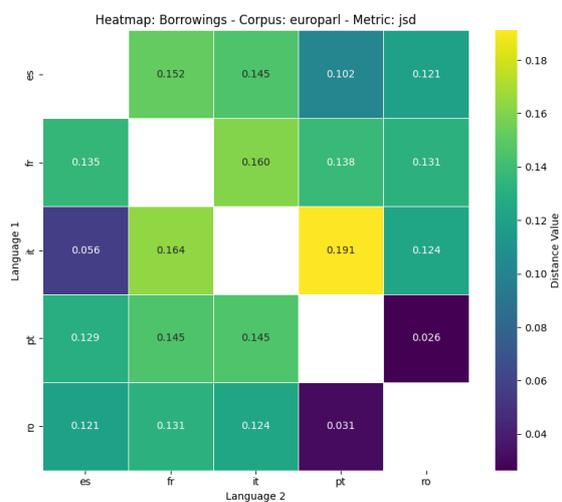


Figure 18: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on Europarl (using JSD)

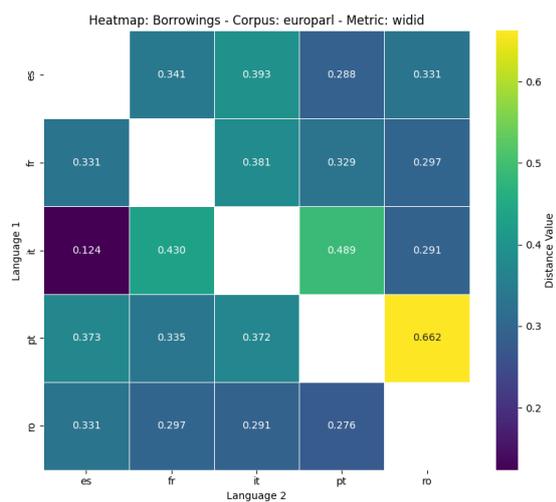


Figure 17: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on Europarl (using widid)

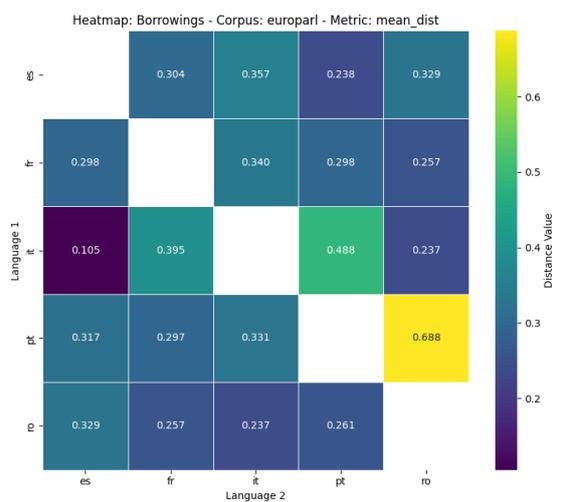


Figure 19: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on Europarl (using mean-dist)

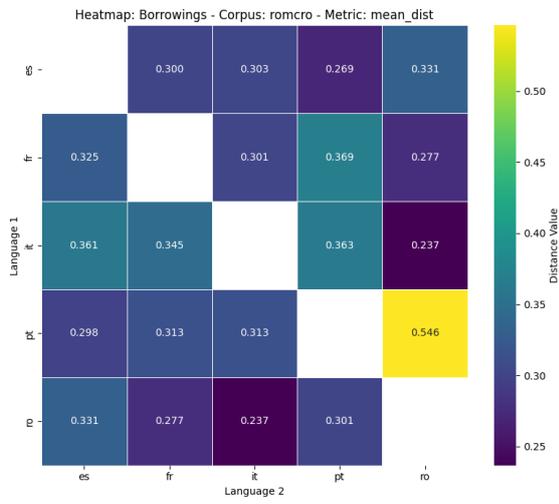


Figure 20: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on RomCro (using mean-dist)

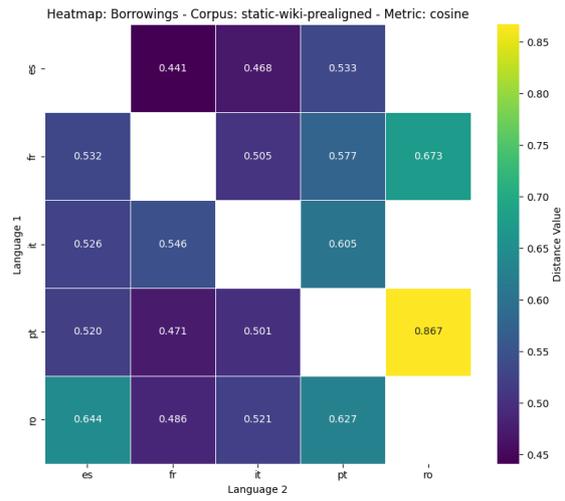


Figure 22: Semantic divergence between borrowings in Romance languages based on static embeddings trained on Wikipedia

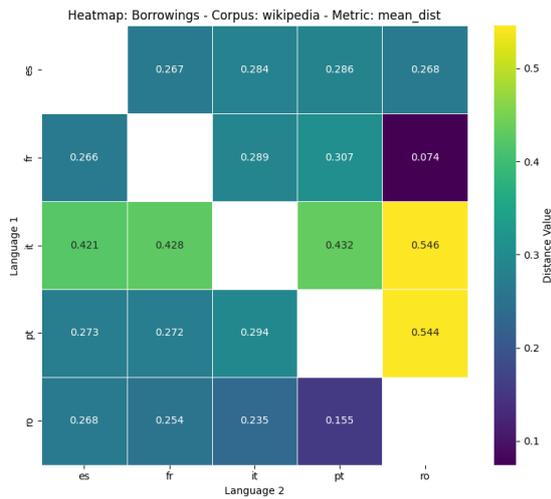


Figure 21: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on Wikipedia (using mean-dist)

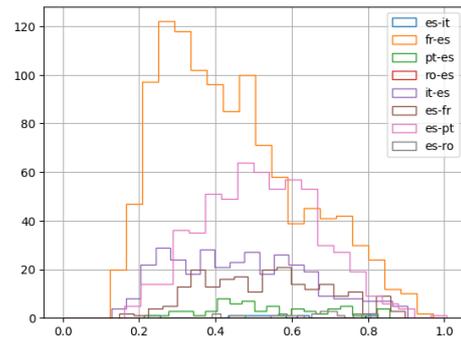


Figure 23: Borrowings distances distribution for Spanish.

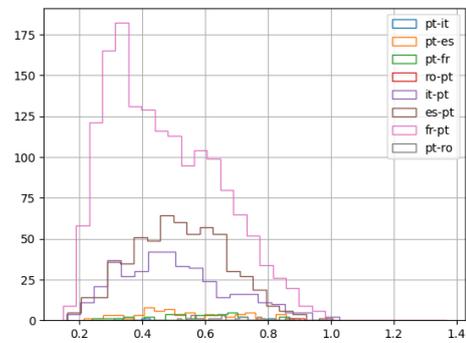


Figure 24: Borrowings distances distribution for Portuguese.

- IT-PT: total: 10421, not in WN: 6479
- IT-RO: total: 2445, not in WN: 1143
- ES-FR: total: 4091, not in WN: 2196
- ES-PT: total: 4018, not in WN: 2131
- ES-RO: total: 5844, not in WN: 3340
- FR-PT: total: 2232, not in WN: 975
- FR-RO: total: 3416, not in WN: 1626
- PT-RO: total: 2545, not in WN: 1280

The coverage of cognates in the corpora used is as follows:

- RO: Total ProtoRom Words: 5522, Found in EuroParl: 3357 (60.79%), Found in Wikipedia: 5248 (95.04%)
- IT: Total ProtoRom Words: 7587, Found in EuroParl: 5576 (73.49%), Found in Wikipedia: 7431 (97.94%)
- ES: Total ProtoRom Words: 6361, Found in EuroParl: 5468 (85.96%), Found in Wikipedia: 6342 (99.70%)
- FR: Total ProtoRom Words: 3991, Found in EuroParl: 3160 (79.18%), Found in Wikipedia: 3952 (99.02%)
- PT: Total ProtoRom Words: 9107, Found in EuroParl: 5851 (64.25%), Found in Wikipedia: 8391 (92.14%)

Threshold-Calibrated Word Sense Disambiguation: Semantic Broadening Without Sense Redistribution in *Schizophrenia*

Naomi Baes^Ψ Nick Haslam^Ψ

^Ψ Melbourne School of Psychological Sciences
The University of Melbourne
n.baes, nhaslam@unimelb.edu.au

Abstract

Polysemous words pose a challenge for computational approaches to language change. We extend a recent hypothesis-driven, prototype-based framework to estimate word sense prevalence in diachronic text corpora and apply it to 109,940 usages of *schizophrenia* drawn from U.S. news media (1985–2025). Our extensions include a contextual dispersion measure (Breadth), robust prototype construction, and human-calibrated prototype-similarity thresholds for conservative sense assignment at scale. Across four decades, distributional semantic change indices commonly used in lexical semantic change detection (LSCD) show significant increases in Breadth and baseline-relative semantic drift (APD), while changes in the central usage prototype (PRT) are influenced by term frequency. In contrast, threshold-calibrated sense assignments reveal stable sense proportions: the psychiatric sense remains dominant, with split-personality and metaphorical senses consistently marginal. Together, these results demonstrate that dispersion- and drift-based LSCD metrics can increase even under stable sense prevalence, indicating that such increases can occur without sense redistribution and primarily reflect broad shifts in usage distributions rather than evidence of polysemization or sense loss. We introduce a threshold-calibrated, prototype-based sense-tracking pipeline that enables conservative sense prevalence estimation at scale and clarifies whether rising distributional LSCD metrics reflect sense redistribution or increasing contextual diversity when historical sense annotation is limited.

Code: [🔗 threshold-calibrated_wsd](#)

1 Introduction

Polysemous words pose a methodological challenge for computational approaches to language change. Lexical semantic change is often studied using distributional methods that quantify semantic

drift from contextualized embeddings, but interpreting these signals can require *sense-aware* modeling that distinguishes between a word’s different contextual usages. For polysemous terms, distributional LSCD metrics are difficult to interpret because they do not directly encode sense distinctions and historical sense annotation is scarce. Therefore, it is often unclear to what extent observed change scores reflect shifts in underlying word senses.

This challenge is commonly framed as *Word Sense Disambiguation* (WSD), the task of computationally identifying which meaning of a word is intended in context (Navigli, 2009). While WSD has a long history in Natural Language Processing, its relevance to historical semantic change research has gained renewed attention in work on lexical semantic change detection (LSCD), driven in part by the availability of contextualized embeddings and sense-aware modeling frameworks (Cassotti et al., 2023; Tang et al., 2023; Periti and Tahmasebi, 2024; Periti and Montanelli, 2024; Aida and Bollegala, 2025). Recent studies have explored diachronic WSD through language model-based classification and targeted retrieval (Beelen et al., 2021; Yadav and Schlechtweg, 2025), unsupervised sense induction with human annotation (Schlechtweg et al., 2025; Goworek et al., 2025), and hypothesis-driven prototype-based retrieval (Cassotti and Tahmasebi, 2025a). Yet, diachronic WSD remains challenging: sense inventories often fail to reflect real-world usage, metaphorical extensions emerge gradually, and annotated historical data are extremely sparse.

The term *schizophrenia* illustrates these challenges. Although originally coined as a psychiatric diagnosis, it has developed multiple meanings in contemporary English, including non-psychiatric and metaphorical uses. Historical dictionaries (e.g., *Oxford English Dictionary*; *Merriam-Webster*) and corpus evidence document several recurrent senses, including a psychiatric sense, a “split personality” sense grounded in a common misconception of the

disorder, and broader metaphorical uses denoting instability or contradiction. These usages imply a pattern of semantic broadening beyond the original psychiatric meaning, especially in public and media discourse. Simultaneously, the psychiatric sense of *schizophrenia* has remained institutionally stable within diagnostic classifications over the past century (Fabiano and Haslam, 2020), making the term a useful test case for examining whether non-psychiatric senses have become more prevalent.

Understanding whether such usages can be reliably distinguished is important for two reasons. First, sense-labeled data provide a foundation for *sense-aware* semantic change detection, a key challenge in computational semantic change detection (Hengchen et al., 2021; Kutuzov et al., 2018). Second, accurate sense decomposition of usage enables more valid measurement of semantic expansion, a phenomenon studied not only in linguistics and natural language processing, but also across the social sciences. For example, psychologists examine “concept creep” (Haslam, 2016), the tendency for harm-related concepts to expand their meanings over time by referring to a broader range of contexts while preserving their definitional core.

Accordingly, the present study introduces a scalable word sense tracking pipeline for diachronic text corpora and applies it to *schizophrenia* in a corpus of U.S. news articles (1985–2025). Using a sense inventory, we examine whether this sense-aware pipeline can reliably distinguish the psychiatric sense of *schizophrenia* from its split-personality and metaphorical uses, and whether changes in distributional semantic change metrics correspond to shifts in sense prevalence. It asks: **(RQ1)** Can a theory-driven sense-aware pipeline distinguish the psychiatric sense of *schizophrenia* from its split-personality and metaphorical uses in a diachronic news corpus?; **(RQ2)** How has the relative distribution of these senses changed?; **(RQ3)** Do LSCD metrics reflect shifts in sense prevalence?

In addressing these questions, the present study contributes (i) a scalable, hypothesis-driven sense tracking pipeline with robust prototypes and human-calibrated thresholds for sense assignment and prevalence estimation; (ii) a large historical human-annotated calibration and evaluation set for *schizophrenia* in U.S. news; and (iii) empirical evidence that increases in dispersion- and drift-based LSCD metrics for polysemous targets do not necessarily reflect changes in sense prevalence, but may instead arise from contextual diversification within

stable senses. In this corpus, observed semantic change reflects contextual diversification within a stable dominant sense rather than prototype displacement or sense replacement.

2 Related Work

Research on diachronic WSD builds on long-standing work in word sense disambiguation (Navigli, 2009) as well as more recent advances in lexical semantic change detection (LSCD). Early LSCD approaches primarily modeled semantic change through distributional shifts in static embeddings (Kutuzov et al., 2018), motivating subsequent *sense-aware* methods aimed at distinguishing distinct and emerging meanings (Hengchen et al., 2021). With the adoption of contextualized and transformer-based representations, more recent studies have demonstrated improved capacity for capturing fine-grained semantic distinctions (Periti and Tahmasebi, 2024; Periti and Montanelli, 2024). However, many of these approaches do not yield interpretable or temporally stable estimates of sense prevalence. Related work has also reframed WSD as a contextual similarity task in WiC-style settings, enabling scalable sense discrimination without full supervision (Cassotti et al., 2023; Yadav and Schlechtweg, 2025).

More recent approaches seek to bridge traditional WSD and unsupervised Word Sense Induction by incorporating expert knowledge, human annotation, and interpretable representations of sense structure (Goworek et al., 2025; Schlechtweg et al., 2025). In particular, Cassotti and Tahmasebi (2025a) propose a hypothesis-driven, generative-prototype framework that supports interpretable tracking of senses over time without relying on induced sense inventories. Nevertheless, their final stage relies on human annotation of sentences most similar to sense prototypes, which constrains scalability in large diachronic text corpora.

Prior content-analytic research has documented non-psychiatric uses of *schizophrenia* in news media, including metaphorical applications denoting inconsistency or contradiction (Duckworth et al., 2003; Chopra and Doody, 2007; Magliano et al., 2011; Cain et al., 2014). Related linguistic work has examined metaphorical uses of *schizophrenia* through concordance analyses (Castaño, 2023). More recent computational studies have focused on Twitter data, showing that references to *schizophrenia* frequently involve non-medical usage and sar-

casm (Joseph et al., 2015; Delanys et al., 2022; Bademli et al., 2023). Yet, social media platforms are relatively recent, and evidence from historical news archives is limited. Consequently, it remains unclear whether non-psychiatric usages of *schizophrenia* have indeed increased in prevalence.

3 Method

3.1 Corpus

We constructed a U.S. news sub-corpus from the U.S. Newsstream Collection (ProQuest Dialog, 2013), which aggregates over one billion English-language articles from more than 1,300 outlets. All articles containing the target term "*schizophrenia*" in the body text were retrieved, deduplicated, and sentence-segmented. The resulting dataset consists of 109,940 cleaned sentences covering the period 1985–2025. Full corpus construction and preprocessing details are provided in Appendix A.

3.2 Measures

3.2.1 Senses

A three-sense inventory was derived from the *Oxford English Dictionary* (OED; 2025), which provides fine-grained, historically attested distinctions for schizophrenia. As detailed in Appendix B, the full four-sense inventory spans literal psychiatric usage, split-personality interpretations, and figurative metaphorical extensions. Lexicographic evidence indicates that, from the post-World War II period (≈ 1945 onward), schizophrenia developed two metaphorical sub-senses: one denoting detachment from reality (3a) and another characterizing internal contradiction or inconsistency (3b). Sub-sense 3a was excluded because its definition is weakly differentiated and overlaps substantially with senses 1 and 3b. Thus, the present study tracks sense prevalence for sense 1 (1908–: the psychiatric condition), sense 2 (1933–: split personality interpretation), and sense 3b (1958–: metaphorical). All three senses are expected to occur in the corpus, as it includes news articles from 1985.

3.2.2 Contextualized Embeddings

Each sentence was encoded using XL-LEXEME (XLL; Cassotti et al., 2023)¹, yielding a single 1024-dimensional *target-conditioned* embedding per sentence. We obtain this usage embedding by

¹<https://huggingface.co/pierluigic/xl-lexeme>. XLM-RoBERTa-large backbone; 24 transformer layers; hidden size 1024; 16 attention heads.

providing XLL with the target token span (start/end indices) for each sentence, producing a pooled representation explicitly conditioned on the marked target rather than a generic sentence embedding. XLL is a bi-encoder with a Siamese (SBERT-style) architecture (Reimers and Gurevych, 2019) fine-tuned on the Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2019) using contrastive learning to increase cosine similarity for sentence pairs expressing the same sense of a target word. This study specifies the target span via start-end indices rather than delimiter tokens, yielding a pooled vector sensitive to the marked target. We adopted XLL because it has demonstrated high sensitivity to induced semantic breadth among sentence-encoding baselines (Baes et al., 2025) and strong WiC performance relative to larger models (Periti and Tahmasebi, 2024). These vectors serve as word-in-context representations of the target term.

3.2.3 Graded Distributional Semantic Change

We quantify lexical semantic change using three complementary measures (see Table 5): a novel Breadth Score (Baes et al., 2024) and two established graded semantic change metrics (Periti and Montanelli, 2024). All measures operate on the same target-conditioned sentence-level embeddings (Section 3.2.2), treated as word-in-context usage representations. Breadth captures within-year contextual dispersion; Average Pairwise Distance (APD) captures baseline-relative distributional divergence between periods and is sensitive to redistribution across usage regions (Giulianelli et al., 2020; Kutuzov and Giulianelli, 2020); and the Prototype Representation Technique (PRT) tracks movement of the central tendency of usage (Martinc et al., 2020; Kutuzov and Giulianelli, 2020). APD and PRT are computed relative to the earliest year (1985) to yield baseline-relative drift scores. To assess robustness to term frequency, we apply a frequency-capped Breadth variant (500 sentences per year) and estimate frequency-controlled regressions including log-transformed annual sentence frequency. Diachronic change is estimated as the regression coefficient over time ($p < .05$), with bootstrap standard errors computed by resampling 500 sentences per year. Full mathematical definitions are provided in Appendix C. For interpretive purposes, we additionally examine year-level associations among the three indices and test whether Breadth statistically accounts for variation in APD and PRT using frequency-controlled regressions.

3.3 Threshold-Calibrated Sense Tracking Pipeline

We build on Cassotti and Tahmasebi’s (2025a) hypothesis-driven framework, treating diachronic WSD as sense-prevalence tracking. In the present study, we extend prototype-based retrieval with a tailored sense inventory for *schizophrenia* and human-calibrated cosine-similarity thresholds for conservative, scalable sense assignment. Figure 1 illustrates the resulting six-stage pipeline.

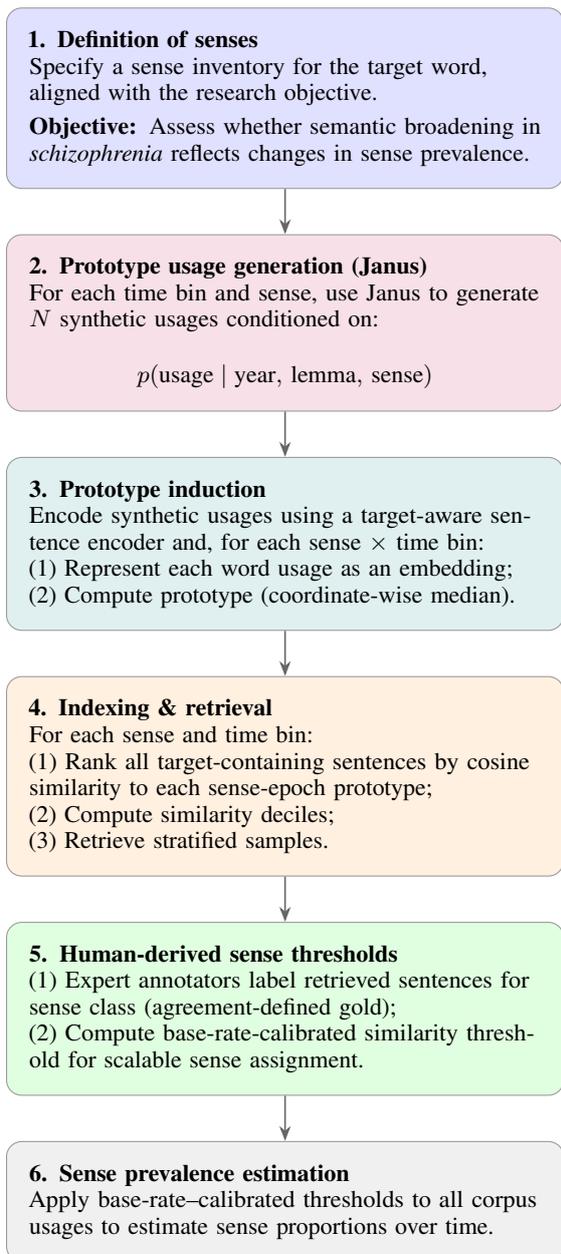


Figure 1: Six-stage threshold-calibrated sense tracking pipeline for hypothesis-driven, prototype-based sense assignment and prevalence tracking in text corpora.

Unlike conventional WSD with fixed inventories or unsupervised word sense induction (e.g., Giulianelli et al., 2020), our approach grounds sense hypotheses in lexicographically aligned generative examples and produces reusable artifacts (prototypes, similarity rankings, and calibrated thresholds) supporting prevalence estimation over time.

(1) Sense inventory. As discussed in Section 3.2.1, we use three target senses of *schizophrenia* from the OED inventory, spanning (i) the literal psychiatric sense, (ii) a split-personality sense, and (iii) a metaphorical extension (contradiction).

(2) Prototype usage generation with Janus. We generated sense-specific synthetic usages using Janus (Cassotti and Tahmasebi, 2025b), a temporally- and sense-conditioned generative model² fine-tuned with QLoRA on 1.2M OED sense-annotated historical usages.³ For each sense and each 5-year time bin (1985–2025), we sampled 500 candidate sentences from: $p(\text{usage} \mid \text{time period}, \text{lemma}, \text{sense})$. Sampling used decoding parameters selected after five human-evaluated pilot rounds (see Appendix D for more detail) to balance determinism and diversity (temperature = 0.6, top- p = 0.7). These settings mitigated common autoregressive artifacts, such as lexical looping and hallucination (Lappin, 2024), while preserving sense-diagnostic lexical cues. Any residual generation noise is addressed downstream via median-based prototype construction.

(3) Prototype induction and quality. All synthetic usages were encoded with XLL into target-conditioned usage embeddings (Section 3.2.2) and filtered for exact duplicates.⁴ From the remaining synthetic pool, we randomly sampled 200 sentences per sense and time bin. Sense prototypes were then computed as coordinate-wise medians of these target-conditioned usage embeddings, aggregating word-in-context sentence representations.

Senses 1 and 2 correspond to mainly literal clinical and lay interpretations of the term, whereas Sense 3b captures metaphorical extensions. Global

²Model: ChangIsKey/llama3-janus (8B parameters). <https://huggingface.co/ChangeIsKey/llama3-janus>.

³Base model: meta-llama/Meta-Llama-3-8B <https://huggingface.co/meta-llama/Meta-Llama-3-8B>, a causal decoder-only transformer (32 layers, 4096 hidden units, 32 attention heads); training data spans 1700–2020. Training format: <year><|t|><lemma><|t|><definition><|s|><usage><|end|>.

⁴Duplicates due to generating 500 sentences per epoch.

summary statistics indicate strong internal coherence across prototypes, with mean intra-sense cosine similarity $\geq .975$ and low variance. Senses 1 and 2 show the highest stability, characterized by low dispersion ($MAD \leq .002$) and relatively small outlier proportions (12–15%). Sense 3b shows greater variability ($MAD = 0.0033$), with higher dispersion and outlier rates (20.3%).

Across senses, embedding norms and centroid magnitudes are highly consistent, suggesting prototype differences are not driven by vector scale. Inter-sense centroid cosine similarity is extremely high for Senses 1–2 (mean ≈ 0.999) but lower for literal versus metaphorical contrasts (Senses 1/2 versus Sense 3b; mean ≈ 0.94 – 0.95), indicating that prototype geometry distinguishes literal from metaphorical usage more strongly than it distinguishes fine-grained literal senses. Because cosine similarities in transformer embedding spaces can be uniformly high, we interpret prototype coherence and separation comparatively (within- versus between-sense). See Appendix D.2 for diagnostics.

(4) Indexing and retrieval. All corpus sentences containing *schizophrenia* were encoded into target-conditioned usage embeddings (Section 3.2.2) and ranked by cosine similarity to each sense prototype. For each sense \times time bin, we computed similarity deciles and drew a stratified sample of 20 sentences per decile (200 per bin), covering the full similarity distribution. Because each embedding is explicitly conditioned on the marked target token, these sentence-level vectors function as contextualized lexical (word-in-context) representations, ensuring representational consistency between prototype induction and retrieval.

(5) Human-derived sense thresholds. To enable scalable yet conservative sense assignment, we calibrated cosine similarity thresholds from expert annotations. For each sense, two annotators applied binary judgments (expresses the target sense or does not) using OED definitions; only unanimous labels were treated as gold. Because minority senses are rare, we used a two-round, stratified design to obtain reliable thresholds under severe class imbalance. In **Round 1**, we sampled uniformly across cosine similarity deciles for each sense prototype and computed *sense purity* (the proportion of unanimously labeled sentences matching the target sense) as a diagnostic check of whether a simple decile-based cutoff was possible. This strategy worked for the dominant psychiatric sense but

yielded too few positives for the split-personality and metaphorical senses (which were underrepresented in the lower similarity deciles). We therefore conducted **Round 2**, enriching annotations in the highest-similarity region (top decile), where true positives for rare senses concentrate. Combining both rounds, we estimated each sense’s corpus *base rate* by weighting similarity regions by their corpus mass, and set one global threshold per sense by selecting the cosine cutoff on the scored candidate set such that the number of retained sentences matched this estimated base rate, yielding conservative thresholds that limit over-assignment under class imbalance. Base-rate estimates and resulting thresholds are reported in Table 3; full estimator details and validation are provided in Appendix E.

(6) Sense prevalence estimation. To estimate sense prevalence, we applied base-rate-calibrated sense thresholds to all scored corpus usages (1985–2025), assigning a sense only when its cosine similarity exceeded the calibrated cutoff. Sentence–prototype similarity distributions show minimal diachronic drift across senses and time bins (Appendix D.3), supporting the use of global thresholds for prevalence estimation. Base rates, estimated from the stratified annotation sample, were used only to set conservative decision thresholds. Prevalence trajectories were then computed on the full corpus as the proportion of usages assigned to each sense after thresholding.

4 Results

Results are presented in stages aligned with the research questions. We first establish semantic change in uses of *schizophrenia* using standard LSCD metrics, motivating the subsequent sense-aware analyses. We then assess sense separability and diachronic sense prevalence (RQ1–RQ2), and finally evaluate whether LSCD trends correspond to shifts in sense prevalence (RQ3).

Evidence for lexical semantic change. We first examined lexical semantic change as captured by standard distributional LSCD metrics in uses of *schizophrenia* between 1985 and 2025 using three measures illustrated in Figure 2. Over this period, the term is used in an increasingly diverse range of contexts, reflected in a significant increase in contextual dispersion (Breadth). Divergence from earlier usage patterns also increases over time (APD), indicating growing distributional semantic drift rel-

ative to the 1985 baseline. In contrast, shifts in the central tendency of usage remain modest (PRT). Together, these patterns indicate increasing contextual dispersion and baseline-relative drift, while the central reference point of usage remains stable.

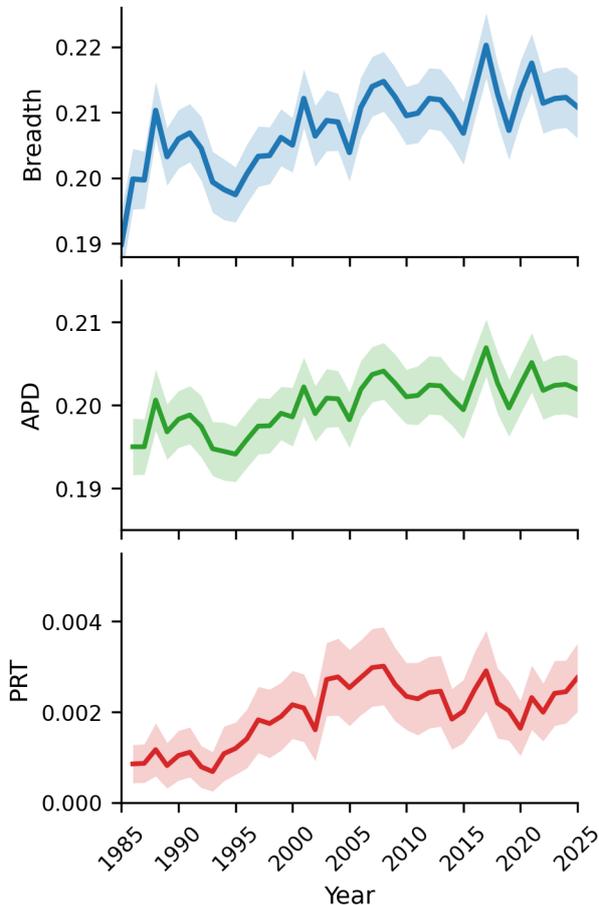


Figure 2: Lexical semantic change indices for *schizophrenia* in U.S. news (1985–2025). (A) Breadth: within-year contextual dispersion, reflecting heterogeneity in usage contexts. (B) APD: average pairwise cosine distance between usages in each year and those in the 1985 reference year, capturing baseline-relative distributional drift. (C) PRT: cosine distance between each year’s centroid and the prototype induced from the 1985 baseline, capturing movement of the central usage pattern without implying sense replacement.

Consistent with these patterns, time-series regression analyses summarized in Table 1 show significant positive temporal trends for Breadth and APD, indicating increasing contextual dispersion and cumulative baseline-relative distributional drift in uses of *schizophrenia*. These trends remain significant when controlling for annual sentence frequency, indicating that the observed changes are not reducible to increases in term prevalence. Results for Breadth remain significant under frequency-capping (500 sentences per year; $\beta = .00035$, CI = [.00024, .00052], $p < .001$,

$R^2 = .55$), further supporting a robust increase in contextual semantic breadth. In contrast, although PRT demonstrates a positive temporal trend in the uncontrolled model, this effect attenuates and becomes non-significant once frequency is controlled.

Metric	β_{Year}	95% CI	p	R^2
Breadth	.00038	[.00024, .00051]	<.001	.55
+	.00021	[.00008, .00035]	.003	.61
APD	.00020	[.00014, .00026]	<.001	.57
+	.00012	[.00005, .00019]	<.001	.62
PRT	.000043	[.000031, .000054]	<.001	.51
+	.000007	[-.000020, .000034]	.595	.72

Table 1: Linear trend estimates for the three semantic change indices for *schizophrenia*. β denotes the slope of the time series regression over years. + = Frequency-controlled models include log annual sentence count as a covariate (HC3 robust standard errors). *Note.* Coefficients (β) are per publication year; one-unit change corresponds to one year, yielding small magnitudes.

Relationships among LSCD measures. To clarify how the distributional indices relate to one another, we examined year-level associations among Breadth, APD, and PRT across 1985–2025. Breadth and APD were almost perfectly correlated (Pearson’s $r = .995$, $p < .001$), indicating that cumulative distributional drift closely tracks increasing contextual dispersion. In contrast, associations involving PRT were substantially weaker (Breadth–PRT: $r = .74$; APD–PRT: $r = .80$, both $p < .001$), suggesting that prototype displacement captures a related but distinct signal. Consistent with this pattern, year-level regressions (Table 2) show that Breadth robustly predicts APD both with and without frequency control, whereas its association with PRT attenuates once frequency is controlled. Results suggest that APD is closely aligned with Breadth, whereas PRT captures a distinct signal that is more sensitive to frequency control.

Outcome	β_{Breadth}	95% CI	p	R^2
APD	.576	[.560, .592]	<.001	.99
+	.537	[.519, .554]	<.001	.99
PRT	.095	[.071, .120]	<.001	.54
+	.038	[.011, .065]	.007	.77

Table 2: Year-level regressions testing whether contextual dispersion (Breadth) predicts baseline-relative drift (APD) and prototype displacement (PRT). Models marked with + are frequency-controlled and include log annual sentence count as a covariate (HC3 robust standard errors).

Sense assignment and prevalence. Addressing RQ1, the threshold-calibrated pipeline enables conservative, scalable sense assignment that reliably distinguishes the psychiatric sense of *schizophrenia* from its split-personality and metaphorical usages, using global, base-rate-calibrated cosine thresholds (Table 3). Consistent with the rarity of the split-personality and metaphorical senses, their calibrated thresholds are substantially higher than for the dominant psychiatric sense, enforcing a precision-first assignment rule under severe class imbalance. Applying these thresholds leaves around 16% of usages below all sense-specific cut-offs (unclassified; Appendix F), which we interpret as low-similarity or ambiguous cases rather than forcing marginal assignments.

Sense	Threshold	Base rate
1 (psychiatric)	0.588	0.800
2 (split-personality)	0.884	0.012
3b (metaphorical)	0.873	0.031

Table 3: Global base-rate-calibrated cosine similarity thresholds by sense.

Addressing RQ2, across five-year bins the relative distribution of *schizophrenia* senses remains stable, as illustrated in Figure 3. When normalized over assigned usages, the psychiatric sense consistently dominates throughout the corpus period (mean proportion = 0.95; range = 0.92–0.96), while the split-personality sense (mean = 0.01; range = 0.01–0.01) and the metaphorical sense (mean = 0.04; range = 0.03–0.07) remain marginal. Temporal variability is low across all senses ($\sigma < 0.014$), indicating stable relative sense distributions over time. Taken together with the LSCD results, the stability of sense prevalence indicates that the observed increases in contextual dispersion and baseline-relative drift do not correspond to shifts in sense prevalence, addressing RQ3.

To aid interpretation, we examined exemplar sentences selected using sentence-level thresholding (Appendix G). For each sense, we inspected both borderline-positive and high-confidence exemplars, illustrating the range of usages admitted by each category. Exemplars for Sense 1 (psychiatric) are qualitatively coherent across confidence levels, whereas exemplars for Senses 2 and 3b more frequently overlap with general or psychiatric mental-health contexts, indicating partial representational overlap. Consequently, small fluctuations in the es-

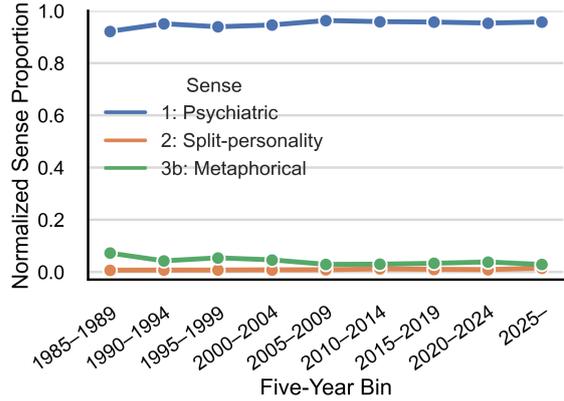


Figure 3: Relative prevalence of *schizophrenia* senses.

timated prevalence of non-dominant senses should be interpreted cautiously. A targeted follow-up analysis directly examined this dissociation and indicates increasing contextual heterogeneity within the dominant psychiatric sense (Appendix H).

5 Discussion

The present study introduced a scalable, hypothesis-driven sense-tracking pipeline for estimating sense prevalence in diachronic text corpora, and applied it to 109,940 U.S. news sentences containing *schizophrenia* (1985–2025). This enables a principled comparison between sense-aware prevalence trajectories and distributional LSCD signals (Breadth, APD, PRT), providing an interpretable test of whether semantic broadening reflects shifts in sense prevalence.

Using this case study, we examined whether increases in distributional semantic change metrics correspond to shifts in sense prevalence across an expert-defined sense inventory. Our results show that (i) threshold-calibrated sense assignment reliably distinguishes psychiatric, split-personality, and metaphorical usages of *schizophrenia* at scale; (ii) the relative prevalence of these senses remains remarkably stable over four decades of U.S. news; and (iii) robust increases in Breadth and APD occur without corresponding shifts in sense prevalence, indicating that these LSCD signals reflect dispersion-driven change within the dominant sense rather than redistribution toward minority senses. Together, these findings demonstrate that commonly used LSCD metrics can increase even when sense prevalence is stable, and should therefore be interpreted as signals of contextual diversification rather than direct evidence of polysemization or sense replacement, unless corroborated

by sense-aware analysis. Because similarity thresholds are calibrated to empirical base rates, the resulting prevalence estimates adopt a precision-first strategy under severe class imbalance, prioritizing high-confidence assignments for rare senses over exhaustive coverage.

These results imply a dissociation between dispersion-based and sense-aware signals of lexical semantic change. Year-level analyses show that Breadth almost perfectly predicts APD, while its association with PRT attenuates once usage frequency is controlled for. This pattern suggests that baseline-relative drift in this news corpus is mainly driven by increasing contextual dispersion rather than movement of a central usage prototype. For polysemous targets, this reflects a possible diagnostic interpretability pitfall: rising LSCD scores may reflect diversification within an existing sense rather than redistribution across senses (shifts in the relative prevalence of distinct senses). In the present study of *schizophrenia*, the observed semantic broadening is consistent with within-sense contextual diversity under a stable sense structure.

Usage-based accounts emphasize that semantic change can occur through gradual expansion in the range of contexts in which a word is used, without requiring redistribution across discrete senses (Traugott and Dasher, 2002; Bybee, 2010). Consistent with the sense-aware analysis and qualitative inspection of high-confidence exemplars, uses of *schizophrenia* increasingly appear across a broader set of institutionally framed contexts (e.g., treatment, legal, and policy reporting). This contextual expansion increases the spread of contextualized representations and elevates dispersion and baseline-relative drift (Breadth, APD), even while the dominant psychiatric sense remains stable. This pattern is also consistent with the term's institutional stability within diagnostic classification systems over the past century (Fabiano and Haslam, 2020), which may constrain large-scale reorganization of its core meaning while permitting diversification in how that meaning is discursively situated. More generally, results align with prior observations that contextualized LSCD measures can assign high change scores in the absence of clear lexicographic sense change, because they may capture contextual variance (and, at times, syntactic redistribution) rather than shifts in dictionary-recordable senses, including sense emergence or loss (Kutuzov et al., 2022). Future work should explicitly decompose global drift into within-sense diversi-

fication versus between-sense redistribution using sense-conditioned dispersion and drift analyses that control for usage volume.

Our results clarify how large-scale, corpus-wide estimates relate to qualitative and content-analytic studies that document substantial non-literal uses of *schizophrenia* in news discourse (Duckworth et al., 2003; Chopra and Doody, 2007; Magliano et al., 2011). These studies show that metaphorical and misconception-driven usages are salient in journalistic writing, but typically rely on purposive sampling or manual coding within comparatively small and temporally restricted samples. For example, in a hand-coded study of Australian print and online news over a one-year period ($N = 630$), only 13% of stories misused *schizophrenia* metaphorically (Cain et al., 2014). In contrast, our estimates are derived from a large historical corpus spanning four decades and employ conservative, high-precision thresholds applied uniformly at scale. Under this design, non-psychiatric uses can be salient without being prevalent: they may cluster in particular outlets, sections, or story types, yet still constitute a small fraction of total occurrences when aggregated across reporting contexts. This highlights a distinction between discursive salience and corpus-wide prevalence. While conservative thresholding may undercount some borderline metaphorical instances, the stability of sense proportions alongside rising dispersion suggests that such undercounting is unlikely to account for the observed increase in Breadth and APD. This interpretation is further supported by comparable content-analytic findings across U.S., UK, Italian, and Australian news media (Duckworth et al., 2003; Chopra and Doody, 2007; Magliano et al., 2011; Cain et al., 2014). Finally, as news discourse is relatively regulated, broader or more stigmatizing metaphorical extensions may be more prevalent in less constrained domains (e.g., social media, everyday language), where editorial norms exert weaker pressure.

The findings also speak to psychological accounts of semantic expansion. From a concept creep (Haslam, 2016) perspective, semantic broadening appears compatible with a largely preserved definitional core: *schizophrenia* remains primarily a psychiatric label, while its usage becomes more contextually flexible, signaling its expanded semantic boundaries. This pattern aligns most closely with what Haslam (2016) terms horizontal semantic expansion, in which a concept's meaning extends through application in a wider range of contexts.

These results have implications for interpreting distributional LSCD metrics. High scores on dispersion- or drift-based measures should not be equated with changes in sense prevalence without complementary sense-aware validation. While Breadth, APD, and related metrics capture shifts in the geometry of usage distributions, such shifts can arise from multiple mechanisms, including within-sense diversification, redistribution between senses, or changes in genre composition. The threshold-calibrated pipeline introduced here provides an interpretable diagnostic for distinguishing among these possibilities by grounding sense hypotheses in lexicographic definitions and estimating prevalence via conservative, human-validated thresholds. Used alongside existing LSCD workflows, this approach enables researchers to better interpret any observed lexical semantic drift. Practically, sense tracking can be treated as a diagnostic tool: when Breadth or APD increase, researchers can examine whether sense proportions also shift over time.

In conclusion, usages of *schizophrenia* in U.S. news between 1985 and 2025 demonstrate clear distributional broadening — marked by increasing contextual dispersion and baseline-relative semantic drift — while the relative prevalence of expert-defined senses remains stable. In particular, the psychiatric sense continues to dominate throughout the study period. This dissociation indicates that semantic broadening can arise through increasing contextual heterogeneity within a stable sense, rather than through redistribution toward minority senses. Consequently, dispersion- and drift-based LSCD signals should not be interpreted as evidence of sense redistribution without complementary sense-aware validation. Concretely, the present study contributes (i) a scalable, hypothesis-driven sense tracking pipeline with human-calibrated thresholds, (ii) empirical evidence that dispersion-based LSCD metrics can rise under stable sense prevalence, and (iii) a sense-aware diagnostic for distinguishing semantic broadening from sense redistribution. More broadly, this study underscores the need to interpret distributional signals of semantic change in light of explicit sense hypotheses, and provides a pipeline for doing so in large diachronic text corpora.

6 Limitations

Several limitations should be noted. First, prevalence estimates for minority senses are conserva-

tive by design. We apply high-precision, base-rate-calibrated similarity thresholds that prioritize precision over recall, leaving approximately 16% of usages unclassified. If these cases disproportionately contain misconception-driven or metaphorical framings, our estimates for the split-personality and metaphorical senses likely underestimate their true frequency. Future work could assess this uncertainty by annotating stratified samples of unclassified cases and by evaluating sensitivity under alternative thresholding schemes. Dynamic thresholds could also be explored if similarity distributions drift over time, potentially inspired by Bayesian approaches (Frermann and Lapata, 2016), as global thresholds may mask gradual boundary shifts.

Second, although our analysis combines dispersion-based LSCD measures with sense-aware validation, it does not explicitly separate within-sense diversification from between-sense redistribution. Prior work shows that contextualized embedding-based change indices can increase due to shifts in contextual variance or syntactic distribution, even in the absence of clear lexicographic sense change (Kutuzov et al., 2022). The divergence between increasing Breadth and APD and stable sense proportions is consistent with growing contextual diversity within senses, but we do not formally test this explanation. Future work should use sense-conditioned dispersion measures to distinguish within-sense diversification from sense redistribution (e.g., the proportion of a word’s sense might shift even if no new sense appears) and trace the source of global distributional drift.

Third, the temporal scope of the corpus is constrained to 1985–2025, despite *schizophrenia* having been coined in 1908. Earlier decades contain substantially fewer occurrences in large-scale news archives, limiting reliable estimation of dispersion, prototype structure, and human-calibrated thresholds. Consequently, the selected diachrony reflects a trade-off between historical coverage and statistical stability. Although the present study controls for usage volume through frequency-controlled regressions and frequency-capped dispersion estimates, frequency remains an important background factor in distributional analyses. A purely frequency-driven account would predict corresponding redistribution across senses or substantial prototype displacement; instead, sense proportions remain stable and prototype movement is modest, suggesting that rising Breadth primarily reflects contextual diversification within a stable

dominant sense. Future work could further formalize frequency–dispersion interactions.

Fourth, while the pipeline employs time-specific sense prototypes to preserve temporal fidelity, our results indicate substantial prototype stability over time for *schizophrenia*. This raises the question of whether time-specific prototypes offer meaningful advantages over time-independent (global) prototypes for concepts with stable prototypical cores. Although time-specific prototypes increase computational cost, a consideration for scalability, they may be necessary for targets undergoing stronger sense drift or temporal reconfiguration. In the present case, however, the observed stability suggests that a lighter pipeline using global prototypes may yield comparable results. Future work should therefore explicitly compare time-specific and time-independent prototypes, assessing quantitative differences in sense assignment, prevalence estimates, and downstream LSCD metrics, as well as qualitative differences in retrieved exemplars, to determine when temporal granularity is justified.

Fifth, our pipeline assumes a discrete and temporally stable sense inventory for the target term, enabling threshold-calibrated WSD at scale. However, the boundaries and granularity of lexical senses are difficult to define operationally, and treating a published sense inventory as canonical necessarily abstracts away deeper theoretical and methodological uncertainties (Tahmasebi et al., 2021). To the extent that sense inventories are underdetermined, any fixed inventory provides only an approximation of semantic structure. Future work might integrate adaptive or probabilistic sense representations that better capture intra-sense heterogeneity and evolving sense boundaries.

Finally, our analysis focuses on a single term in a single national news corpus, which limits generalizability and may reflect genre-specific editorial norms rather than broader patterns of semantic change across discourse domains (e.g., podcasts, scientific writing, or everyday language). The observed pattern — stable sense proportions alongside increasing contextual dispersion (particularly in *schizophrenia*’s clinical sense) — is compatible with *determinologization* (Gorokhova, 2020), the linguistic process whereby specialized, technical, or scientific terms diffuse into broader public discourse thereby losing their strict, context-independent definitions to become part of the general vocabulary, while retaining a stable definitional core. Because the present study focuses on dis-

tributional and sense-prevalence dynamics rather than socioterminological boundary shifts, we do not model determinologization directly. Future work should apply this framework to additional polysemous terms (e.g., *intelligence*, *light*, *technology*), other harm-related concepts with multiple dictionary senses (e.g., *trauma*, *depression*), and less regulated discourse domains (e.g., social media), where metaphorical extension and terminological drift may be more prevalent and editorial constraints weaker. Words with more obviously disambiguated senses (e.g., *bank*) may show greater success using this threshold-calibrated word sense tracking method. Integrating formal determinologization frameworks with sense-aware dispersion diagnostics remains an important direction for future research.

Acknowledgments

We thank Change is Key! for hosting the first author at the University of Gothenburg, where early methodological exploration of sense induction and contextualized embeddings informed this work. We are grateful to Nina Tahmasebi for valuable methodological discussions, and to Pierluigi Casotti for suggesting the use of the hypothesis-driven framework for word sense disambiguation. We also thank Haim Dubossarsky for mentorship on semantic change metrics and for constructive discussions clarifying the distinctions between Breadth, APD, and PRT, as well as Ekaterina Vylomova for her guidance in the development of the Breadth score, introduced in the original SIBling framework paper, and for her ongoing academic mentorship.

The majority of analyses were run in TDM Studio (U.S. Newsstream ProQuest; export restrictions apply) on an 8-vCPU Intel Xeon Platinum 8259CL @ 2.50 GHz system (4 physical cores / 8 threads, ~30 GB RAM). Transformer-based models were executed on a GPU-backed Amazon SageMaker instance (NVIDIA Tesla T4, ~16 GB VRAM; ~15 GiB usable) provisioned via TDM Studio.

Janus inference was run on the University of Melbourne’s Spartan HPC (Research Computing Services; Lafayette et al., 2016) via the `gpu-a100` partition; jobs were allocated 8 CPU threads with an NVIDIA A100 (80 GB VRAM).

This research was supported by the Commonwealth through an Australian Government Research Training Program Scholarship (<https://doi.org/10.82133/C42F-K220>), and funded in

part by Australian Research Council Discovery Project DP250102690.

References

- Taichi Aida and Danushka Bollegala. 2025. Investigating the contextualised word embedding dimensions specified for contextual and temporal semantic changes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1413–1437, Abu Dhabi, UAE. Association for Computational Linguistics.
- K. Bademli, A. Kaya Kılıç, and M. Kayakuş. 2023. Using twitter to assess stigma to schizophrenia and psychosis: A qualitative study. *Turkish Journal of Psychiatry*, 34(3):154–161.
- Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2024. A multidimensional framework for evaluating lexical semantic change with social science applications. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1390–1415, Bangkok, Thailand. Association for Computational Linguistics.
- Naomi Baes, Raphael Merx, Nick Haslam, Ekaterina Vylomova, and Haim Dubossarsky. 2025. LSC-eval: A general framework to evaluate methods for assessing dimensions of lexical semantic change using LLM-generated synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10905–10939, Vienna, Austria. Association for Computational Linguistics.
- Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761, Online. Association for Computational Linguistics.
- Joan L. Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press.
- Belinda Cain, Roseanne Currie, Eleanor Danks, Fiona Du, Erica Hodgson, Jennifer May, Kirsty O’Loughlen, Yen Phan, Jennifer Powter, Nayab Rizwan, Shazmi Shahim, Dominique Simsion, Steve Loughnan, and Nick Haslam. 2014. “schizophrenia” in the australian print and online news media. *Psychosis*, 6(2):97–106.
- Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.
- Pierluigi Cassotti and Nina Tahmasebi. 2025a. A hypothesis-driven framework for detecting lexical semantic change. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 177–185, Cagliari, Italy. CEUR Workshop Proceedings.
- Pierluigi Cassotti and Nina Tahmasebi. 2025b. Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics*, 13:690–708.
- E. Castaño. 2023. What is in a word? an exploration of the metaphorical use of schizophrenia in general american english. *Lingua*, 294:103596.
- Anju K. Chopra and Gillian A. Doody. 2007. Schizophrenia, an illness and a metaphor: Analysis of the use of the term “schizophrenia” in UK national newspapers. *Journal of the Royal Society of Medicine*, 100(9):423–426.
- Sarah Delanys, Farah Benamara, Véronique Moriceau, François Olivier, and Josiane Mothe. 2022. Psychiatry on twitter: Content analysis of the use of psychiatric terms in french. *JMIR Formative Research*, 6(2):e18539.
- Kenneth Duckworth, John H Halpern, Russell K Schutt, and Christopher Gillespie. 2003. Use of schizophrenia as a metaphor in us newspapers. *Psychiatric services*, 54(10):1402–1404.
- Fabian Fabiano and Nick Haslam. 2020. Diagnostic inflation in the dsm: A meta-analysis of changes in the stringency of psychiatric diagnosis from dsm-iii to dsm-5. *Clinical Psychology Review*, 80:101889.
- Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
- Natalia V Gorokhova. 2020. Determinologization and transterminologization processes in modern oil and gas discourse. In *European Proceedings of Social and Behavioural Sciences EpSBS*, pages 329–335.
- Roksana Goworek, Harpal Singh Karlcut, Hamza Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Satkirti Purighella, Paridhi Gupta, Muhinyia Ndegwa, Bao Khanh Tran, and Haim Dubossarsky. 2025. SenWiCh: Sense-annotation of low-resource languages for WiC using hybrid methods. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 61–74, Vienna, Austria. Association for Computational Linguistics.

- Nick Haslam. 2016. [Concept creep: Psychology’s expanding concepts of harm and pathology](#). *Psychological Inquiry*, 27(1):1–17.
- Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. [Challenges for computational lexical semantic change](#). In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Computational Approaches to Semantic Change*, pages 341–372. Language Science Press, Berlin.
- Andrew J. Joseph, Neeraj Tandon, Lawrence H. Yang, Kenneth Duckworth, John Torous, Larry J. Seidman, and Matcheri S. Keshavan. 2015. [#schizophrenia: Use and misuse on twitter](#). *Schizophrenia Research*, 165(2–3):111–115.
- Andrey Kutuzov and Mario Giulianelli. 2020. [UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. [Diachronic word embeddings and semantic shifts: a survey](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. [Contextualized embeddings for semantic change detection: Lessons learned](#). *Northern European Journal of Language Technology*, 8.
- Shalom Lappin. 2024. [Assessing the strengths and weaknesses of large language models](#). *Journal of Logic, Language and Information*, 33(1):9–20.
- Lorenza Magliano, John Read, and Riccardo Marassi. 2011. [Metaphoric and non-metaphoric use of the term “schizophrenia” in Italian newspapers](#). *Social Psychiatry and Psychiatric Epidemiology*, 46(10):1019–1025.
- Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. [Leveraging contextual embeddings for detecting diachronic semantic shift](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.
- Merriam-Webster. 2025. [Schizophrenia](#). Merriam-Webster.com Dictionary.
- Roberto Navigli. 2009. [Word sense disambiguation: A survey](#). *ACM Comput. Surv.*, 41(2):10:1–10:69.
- Oxford English Dictionary. 2025. [Oxford english dictionary](#). Accessed 13 December 2025.
- Francesco Periti and Stefano Montanelli. 2024. [Lexical semantic change through large language models: a survey](#). *ACM Comput. Surv.*, 56(11).
- Francesco Periti and Nina Tahmasebi. 2024. [A systematic comparison of contextualized word embeddings for lexical semantic change](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. [WiC: The word-in-context dataset for evaluating context-sensitive meaning representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.
- ProQuest Dialog. 2013. [How do i remove duplicate records?](#) <https://pq-static-content.proquest.com/collateral/media2/documents/pqd-hdi-remove-duplicate-records.pdf>. ProQuest Support Center.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Dominik Schlechtweg, Frank D Zamora-Reina, Felipe Bravo-Marquez, and Nikolay Arefyev. 2025. [Sense through time: Diachronic word sense annotations for word sense induction and lexical semantic change detection](#). *Language Resources and Evaluation*, 59(2):1431–1465.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. [Survey of computational approaches to lexical semantic change detection](#). In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, pages 1–91. Language Science Press.
- Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. [Can word sense distribution detect semantic changes of words?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.
- TDM Studio. 2023. [Tdm studio](#). <https://tdmstudio.proquest.com>. ProQuest, part of Clarivate. Ann Arbor, MI, USA. Accessed: 2023–2025.
- Elizabeth Closs Traugott and Richard B. Dasher. 2002. *Regularity in Semantic Change*. Cambridge University Press, Cambridge.
- Sachin Yadav and Dominik Schlechtweg. 2025. [Xl-durel: Finetuning sentence transformers for ordinal word-in-context classification](#). *arXiv preprint arXiv:2507.14578*.

A Corpus Construction and Preprocessing Details

A.1 Data Source

Data were sourced from the U.S. Newsstream Collection using TDM Studio (TDM Studio, 2023), which provides access to ProQuest databases and over one billion English-language news articles from over 1,300 U.S. national and regional outlets, including major dailies (e.g., The New York Times, The Washington Post, Los Angeles Times) and six regional collections (e.g., Midwest, Southeast). We retrieved all English-language news articles in which the target *schizophrenia* appeared in the body text (excluding title- or abstract-only hits). Articles spanned 176 publishers, 49 provinces, and eight source types (Audio & Video Works; Blogs, Podcasts & Websites; Magazines; Newspapers; Other Sources; Reports; Trade Journals; Wire Feeds).

A.2 Deduplication

ProQuest text contains a high proportion of near-duplicate articles (ProQuest Dialog, 2013). We applied a shingling-based deduplication pipeline, following Pietsch et al. (under review.), using 5-gram character shingles, MinHash, and locality-sensitive hashing. Candidate pairs with estimated similarity > 0.4 were re-evaluated using exact Jaccard similarity; pairs with Jaccard > 0.6 were marked as duplicates. Duplicate articles were grouped into clusters, and only the longest article in each cluster was retained. A validation test on a 1,000-article random sample yielded a median Jaccard similarity of 0, confirming that the remaining corpus contained distinct texts.

A.3 Sentence Extraction

Sentences containing the target term were extracted using spaCy’s rule-based sentencizer. Articles of length $\leq 12,000$ characters were fully segmented into sentences; longer articles were processed by identifying case-insensitive whole-word matches of the target and extracting a ± 500 -character window around each match. Extracted sentences were capped at 600 characters; sentences exceeding this limit were automatically re-windowed using a ± 260 -character span, expanded to the nearest whitespace boundary. Post-processing removed URL fragments, boilerplate text (e.g., strings following “Available at”), fragments with fewer than three tokens, short title or link stubs (≤ 8 words), and table-of-contents-like metadata lines (e.g., heavy section numbering, bullet runs, or strings dominated by digits).

A.4 Final Corpus

For final analyses, years were restricted from 1985-2025 to ensure adequate sampling at time point 1 (375 sentences), yielding a final dataset containing 109,940 cleaned sentences with *schizophrenia* from 70,993 articles. Descriptives entailed sentence length (mean: 183 chars, median: 169 chars; range: 20-1,252 chars), words per sentence (mean: 27; median: 25), article length (mean: 6,083 chars, median: 4,411 chars, range: 74 to 39,551 chars), words per article (mean: 999, median: 724). Figure 4 shows annual counts of sentences mentioning schizophrenia in U.S. news articles.

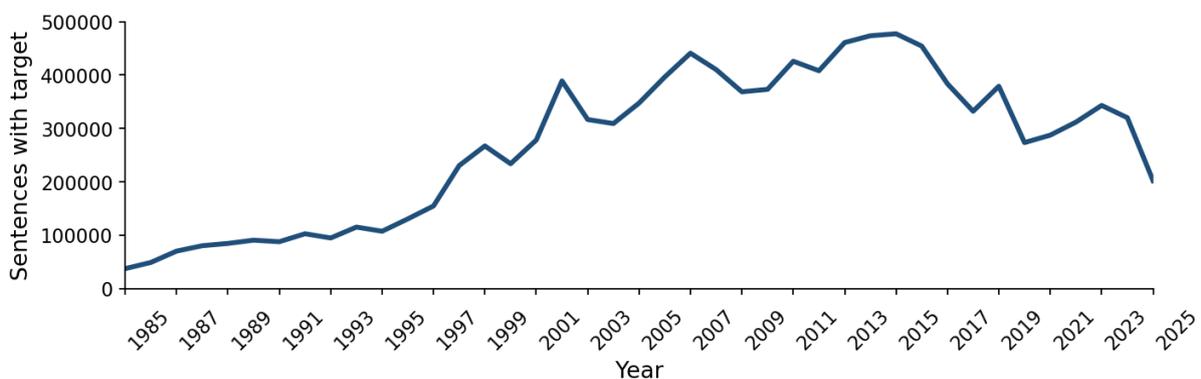


Figure 4: Count of sentences mentioning schizophrenia in the U.S. Newsstream corpus.

B Oxford English Dictionary Senses of Schizophrenia

Note. Sense 3a was excluded because its meaning is poorly differentiated and overlaps with senses 1 and 3b. Janus also struggled to generate reliable examples of this sense, reinforcing its conceptual instability. We therefore remove it from main analyses.

Sense	Definition	Example Usage
1. (1908-)	<i>Psychiatry and Psychology.</i> A mental health condition often having a serious impact upon personal, interpersonal, and occupational functioning, of which typical features are the occurrence of hallucinations and delusions, eccentric speech and behaviour, and diminished emotional expression and purposeful activity. Also: a type or instance of this; any of a spectrum of conditions sharing features of this.	“After all these years of living with schizophrenia, addiction, and drug-induced Parkinsonism, my mother has also been diagnosed with emphysema, hoarding disorder, and several other illnesses.” (J. Díaz, <i>Ordinary Girls</i> iv. 282, 2019)
2. (1933-)	The condition of one individual having, or being supposed to have, two or more distinct personalities between which the individual switches.	“The Brothers’ latest film, <i>Me, Myself and Irene</i> , a romantic comedy about schizophrenia starring Jim Carrey.” (<i>Independent on Sunday</i> , 16 April, Review section, 55/4, 2000)
3a.* (1945-)	Detachment from reality; a sense of alienation from one’s circumstances or environment.	“Colonialism invented . . . a people exiled from their communities, dislocated and suffering from deep alienation: cultural schizophrenia.” (<i>Zimbabwe Independent</i> (Nexis), 21 June, 2019)
3b.* (1958-)	A mentality or approach characterized by inconsistent or contradictory elements.	“Wray knows how to induce and then manage a kind of epistemological schizophrenia in the reader, whereby we can inhabit Lowboy’s groundless visions and still glimpse the ground they negate.” (<i>New Yorker</i> , 30 March, 70/2, 2009)

Table 4: Dictionary senses of *schizophrenia* with example usages.

Note. * = *Figurative*: “Usage without reference to a diagnosed mental health condition is now sometimes avoided as potentially *offensive*.”

C Graded Semantic Change Detection Metrics

This appendix formally defines the graded semantic change metrics used in the present study. Table 5 provides an intuitive, verbal description of what each metric captures conceptually, while the equations below it give their corresponding mathematical formalizations based on cosine distances between contextualized (word-in-context) embeddings. All three measures are cosine-distance-based and are bounded in the interval $[0, 1]$, facilitating direct comparison across metrics.

Form-based Measure	What it Captures
Breadth Score (Within-Year Dispersion)	Computed as the average pairwise cosine distance between all distinct pairs of sentence embeddings within a given year. Higher values indicate that the term is used across a <i>wider and more heterogeneous set of contexts</i> , reflecting greater contextual dispersion within a time period $[0, 1]$.
APD (Average Pairwise Distance)	Computed as the average cosine distance between contextual embeddings drawn from two different time periods. APD captures <i>baseline-relative distributional divergence</i> , reflecting how the overall configuration of usages at time t_2 differs from that at t_1 . Higher values indicate greater separation between the two usage distributions in embedding space, but APD alone does not distinguish between redistribution among existing senses and increasing contextual heterogeneity within a stable sense inventory $[0, 1]$.
PRT (Prototype Representation Technique)	Calculated as the cosine distance between centroid (prototype) embeddings of a word’s contextual usages in two time periods. PRT captures movement of the <i>central tendency of usage</i> in embedding space; higher values indicate greater displacement of the dominant usage pattern, without implying the emergence or replacement of discrete senses $[0, 1]$.

Table 5: Form-based graded semantic change metrics used to quantify distributional semantic broadening.

Formal definitions. Let $\Phi^{t_1} = \{a_1, \dots, a_n\}$ and $\Phi^{t_2} = \{b_1, \dots, b_m\}$ denote the contextual embeddings for a target term in periods t_1 and t_2 , respectively, and let $d(\cdot, \cdot)$ be cosine distance.

Breadth Score (within-year).

$$\text{Breadth}(t) = \frac{1}{\binom{|\Phi^t|}{2}} \sum_{a < b} d(a, b)$$

Average Pairwise Distance (APD).

$$\text{APD}(t_1, t_2) = \frac{1}{|\Phi^{t_1}| |\Phi^{t_2}|} \sum_{a \in \Phi^{t_1}} \sum_{b \in \Phi^{t_2}} d(a, b)$$

Prototype Representation Technique (PRT).

$$\text{PRT}(t_1, t_2) = d(\mu_1, \mu_2), \quad \mu_i = \frac{1}{|\Phi^{t_i}|} \sum_{x \in \Phi^{t_i}} x$$

Note. In the present study, centroids are computed by averaging contextual embeddings and then ℓ_2 -normalizing the resulting vectors prior to distance computation.

D Evaluation of Janus-Generated Diachronic Sense Prototypes

Table 6 summarizes the five iterative rounds of decoding experiments conducted to select hyperparameters for synthetic usage generation. Across rounds, we varied temperature and top- p values to balance semantic accuracy (faithfulness to the intended sense), grammaticality, and contextual diversity. In each round, we sampled at least 10 candidate usages per sense \times period and manually inspected the outputs (with additional inspection triggered by observed failures). Round 1 showed clear hallucinations and excessive repetition, while Rounds 2 and 3 remained insufficiently constrained, with frequent grammatical errors, sense drift, and over-linking to the target lemma. Round 4 produced the most coherent psychiatric, split-personality, and metaphorical usages, with only minor issues (generating examples from national-level politics for sense 3b due to training data, limited contextual variability due to conservative parameters, and possible sense confusion - overcome at the human evaluation stage). Round 5 proved overly conservative, yielding short, templated sentences with reduced lexical diversity. Overall, Round 4 provided the strongest balance of sense fidelity, grammatical clarity, and contextual variability. We therefore adopted temperature = 0.6 and top- p = 0.7 as the final decoding configuration for all subsequent experiments.

Round	Hyperparameters	Main Issues Observed	Representative Examples
1	temp = 1.0 top- p = 0.9 max_new = 50	Hallucinations; repeated target in the same sentence; incoherent continuations (mainly repetitions); Sense 2 usages not representing sense (20 statements do not contain anything about split personality: lines 201-320)	“The lion has schizophrenia.” (<i>nonsensical: only humans can have schizophrenia</i>) “The psychoses were divided into two groups: schizophrenia and schizophrenia.” “The schizophrenia of schizophrenia is not the normal state of the human psyche.” (<i>repetition of the target term</i>)
2	temp = 0.8 top- p = 0.9 max_new = 50	Insufficient constraint; difficulty generating distinct sense types; weak control over registry.	“The schizophrenias were classified into catatonic, hebephrenic and paranoid.” (<i>good</i>) <i>But also:</i> “He has schizophrenia and he has had multiple episodes of coma.” (<i>nonsensical</i>) “A 40-year-old man with schizophrenia and schizophrenia was arrested...” (<i>repetition of the target term</i>)
3	temp = 0.8 top- p = 0.6 max_new = 50	Grammar errors; extreme repetition; sense drift; over-linking to target lemma; model adds extra mental disorders; inability to maintain split-personality sense.	“The psychoses were divided into two groups: schizophrenia and schizophrenia.” (<i>Medical sense, 1990</i>) “A man with schizophrenia and schizophrenia was arrested...” (<i>Medical sense, 2000</i>) “It’s not schizophrenia, but I’m very, very, very, very...” (<i>Split-personality sense, 2000:</i>) “The schizophrenia of the individual is a product of the schizophrenia of the culture.” (<i>Medical sense, 2015–2020:</i>)
4	temp = 0.6 top-p = 0.7 seed = 42 max_new = 50	Best balance of diversity and accuracy; clean psychiatric, split-personality, and figurative/systemic usages; minor awkwardness remains but acceptable.	“Schizophrenia affects about one in a hundred people and often involves hallucinations.” “He felt as if two selves lived inside him, switching without warning.” “The schizophrenia of American politics is its simultaneous demand for small government and high security.”
5	temp = 0.5 top- p = 0.8 seed = 42 max_new = 50	Too conservative; low lexical diversity; templated responses; repetitive surface forms.	“Schizophrenia is a mental disorder affecting thoughts, feelings and behaviour.” (<i>appears repeatedly</i>) “A divided self switches between two personalities.” (<i>repeated across periods</i>) <i>Figurative sense is too generic</i> (“the schizophrenia of the system...”).

Table 6: Summary of hyperparameter rounds for synthetic usage generation.

Note. Examples are shortened for readability. Reported issues reflect manual inspection of generated usages across senses and time bins.

D.1 Illustrative Synthetic Usages

Representative prototype-reflective synthetic sentences are shown in Tables 7–9. These examples illustrate the semantic core captured by each prototype, as reflected in the highest-similarity synthetic usage retrieved for each sense and time period. The examples give an indication of the prototypical sense for psychiatric (Sense 1), split-personality (Sense 2), and metaphorical (Sense 3b). Janus decoding parameters are from round 4: temperature = 0.6, top- p = 0.7, max new tokens = 50.

Year	Sim.	Prototype-reflective example
1985	0.9983	“The schizophrenia is not a single condition. It is a family of conditions.”
1995	0.9985	“The schizophrenia is a group of disorders in which the patient has lost touch with reality.”
2000	0.9987	“A schizophrenia is a brain disorder that affects the normal thought and emotional processes.”
2005	0.9981	“A schizophrenia is a brain disorder that affects about one in 100 people.”
2010	0.9987	“A schizophrenia is a brain disorder that affects about one in 100 people.”
2015	0.9986	“A schizophrenia is a brain disorder that affects how you think, behave, and feel.”
2020	0.9982	“A schizophrenia is a brain illness that affects how you think, behave and feel.”
2025	0.9985	“A schizophrenia is a brain disorder that affects the way you think, feel and behave.”

Table 7: Nearest synthetic sentence (of 200) to prototypes for Sense 1 (psychiatric).

Note. Year denotes the starting year of each five-year interval in the corpus period (1985-2025).

Year	Sim.	Prototype-reflective example
1985	0.9985	“The idea of schizophrenia, or multiple personalities, is not new.”
1990	0.9982	“Schizophrenia is a rare condition in which two personalities coexist in the same body.”
1995	0.9987	“The idea of schizophrenia, or multiple personality, has been around for a long time.”
2000	0.9986	“The idea of schizophrenia, or multiple personality, has been around for a long time.”
2005	0.9986	“The schizophrenia theory is based on the idea that there are two distinct personalities within one person.”
2010	0.9984	“The schizophrenia theory is based on the idea that the person has two personalities.”
2015	0.9983	“A schizophrenia sufferer is supposed to have two personalities.”
2020	0.9981	“I’m not 100% sure that I’m not actually suffering from schizophrenia.”
2025	0.9983	“The schizophrenia is the result of the person’s own self-deception.”

Table 8: Nearest synthetic sentence (of 200) to prototypes for Sense 2 (split-personality).

Note. Year denotes the starting year of each five-year interval in the corpus period (1985-2025).

Year	Sim.	Prototype-reflective example
1985	0.9981	“The schizophrenia of the American mind is manifest in its attitude toward the United Nations.”
1990	0.9978	“The schizophrenia of the American psyche is expressed in the fact that we are the most religious people in the world.”
1995	0.9986	“The schizophrenia of American politics is seen in the contrast between the real and the ideal.”
2000	0.9983	“The schizophrenia of the American mind is most clearly exemplified in the field of foreign policy.”
2005	0.9986	“The schizophrenia of American politics is that we’re a nation of people who are deeply concerned about the fate of other nations.”
2010	0.9982	“The schizophrenia of the American political scene is most evident in the field of foreign policy.”
2015	0.9982	“The schizophrenia of American politics is that we are a nation of immigrants, but we have no policy to deal with them.”
2020	0.9986	“The schizophrenia of American politics today is that we’re both the greatest nation in history and the worst.”
2025	0.9981	“The schizophrenia of the new Americanism is best seen in its attitude toward the United Nations.”

Table 9: Nearest synthetic sentence (of 200) to prototypes for Sense 3b (metaphorical/figurative contradiction).

Note. Year denotes the starting year of each five-year interval in the corpus period (1985-2025).

D.2 Quality Metrics of Sense Prototypes

To assess whether the Janus-generated exemplars produced coherent and sense-distinct contextual clusters, we computed prototype-quality diagnostics for all XL-LEXEME centroids across senses and time periods. Table 10 summarizes global diagnostics, while Figure 5 reports sense-specific diagnostics, including prototype compactness, temporal stability, intra-group similarity, and outlier rates. These metrics evaluate the internal structure of each sense cluster rather than the semantic correctness of the sense itself, and therefore complement the human-validation checks above. Notably, Sense 3a (“detachment / alienation”)

demonstrates acceptable prototype structure across multiple diagnostics, but is excluded from downstream analyses due to its lexicographically diffuse definition and strong overlap with Sense 3b.

Metric	Sense 1	Sense 2	Sense 3a	Sense 3b
Number of embeddings	1,800	1,800	1,800	1,800
Number of 5-year bins	9	9	9	9
Sense centroid norm	30.88	30.89	30.81	30.90
Embedding norm mean	31.07	31.08	31.13	31.15
Embedding norm std	0.075	0.086	0.067	0.042
Intra-sense similarity mean	0.986	0.985	0.976	0.975
Intra-sense similarity std	0.010	0.012	0.016	0.022
Dispersion median	0.0054	0.0051	0.0098	0.0073
Dispersion MAD	0.0019	0.0020	0.0038	0.0033
Outlier proportion	12.2%	15.1%	12.3%	20.3%
Stability score	0.0061	0.0059	0.0110	0.0087

Table 10: Global quality diagnostics for XL-LEXEME sense centroids. Senses 1 and 2 show the highest stability with low dispersion and outlier rates, while Sense 3b exhibits the highest variability and Sense 3a shows elevated dispersion. All senses maintain strong internal coherence with intra-sense similarity ≥ 0.975 .

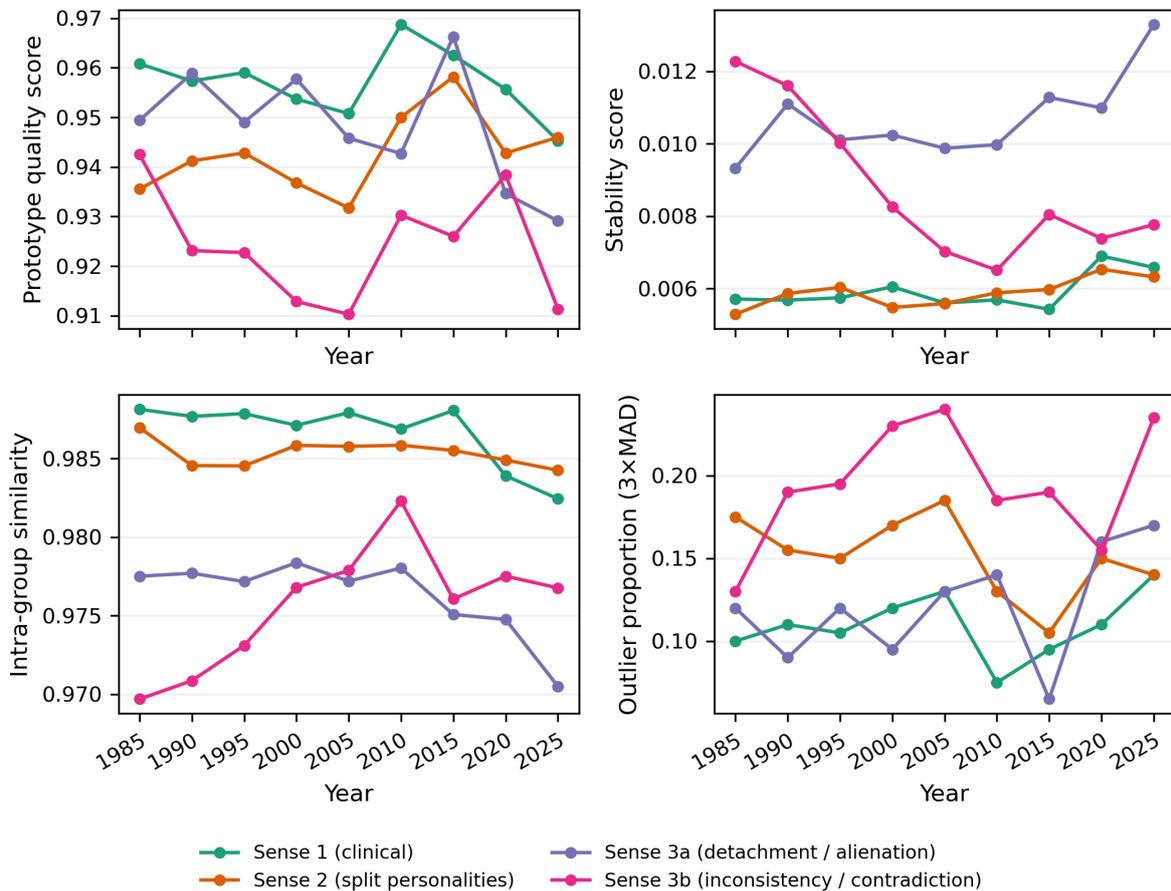


Figure 5: Local prototype-quality diagnostics for XL-LEXEME centroid embeddings across senses and years. (A) Prototype quality; (B) temporal stability; (C) intra-group similarity; and (D) outlier proportion ($3 \times \text{MAD}$). Higher quality and intra-group similarity reflect more compact and coherent sense prototypes, while lower stability scores and fewer outliers indicate temporal consistency. Sense 3a exhibits weaker structural coherence across metrics, consistent with its broad and diffuse definition.

Note. Year denotes the starting year of each five-year interval in the corpus period (1985-2025).

Because cosine similarities in transformer embedding spaces can be uniformly high due to anisotropy, we also computed cosine similarity between centroids of different senses within the same time period (Table 11). Inter-sense centroid similarity is high, consistent with anisotropy in contextual embedding spaces. Nevertheless, centroid similarity is lowest for literal vs. metaphorical contrasts (Senses 1/2 vs

Sense 3b), supporting meaningful separation along the literal–figurative axis; fine-grained distinctions among literal senses are less separable at the centroid level (Table 11).

Sense pair	Mean cosine similarity
Sense 1 – Sense 2	0.999
Sense 1 – Sense 3a	0.989
Sense 2 – Sense 3a	0.992
Sense 3a – Sense 3b	0.977
Sense 1 – Sense 3b	0.942
Sense 2 – Sense 3b	0.950

Table 11: Mean inter-sense centroid cosine similarity across time periods. Values reflect known anisotropy in contextualised embedding spaces but show systematic separation between senses, particularly between literal and metaphorical clusters.

D.3 Descriptive Statistics for Prototype–Sentence Cosine Similarity

Table 12 summarizes cosine similarity statistics between corpus sentences and their XL-LEXEME sense prototypes by each five-year period. Across all senses and epochs, similarity distributions are highly stable: mean and median values cluster tightly around 0.69–0.73, dispersion is low, and both minimum and maximum values fall within a constrained range. Despite large changes in corpus size over time, the distribution of sentence–prototype similarity demonstrates minimal diachronic drift.

Year	Sense	<i>n</i>	Mean	SD	Min	Median	Max
1985	1	372	0.71	0.13	0.24	0.74	0.90
1990	1	895	0.69	0.13	0.19	0.72	0.90
1995	1	1023	0.71	0.13	0.23	0.73	0.90
2000	1	2179	0.69	0.13	0.19	0.72	0.91
2005	1	3104	0.70	0.13	0.15	0.73	0.91
2010	1	3358	0.69	0.13	0.22	0.72	0.91
2015	1	4325	0.70	0.13	0.11	0.73	0.91
2020	1	2468	0.70	0.13	0.22	0.73	0.93
2025	1	1820	0.70	0.13	0.18	0.73	0.93
1985	2	372	0.72	0.12	0.25	0.75	0.90
1990	2	895	0.70	0.13	0.19	0.73	0.90
1995	2	1023	0.71	0.13	0.24	0.74	0.91
2000	2	2179	0.70	0.13	0.19	0.72	0.91
2005	2	3104	0.70	0.13	0.16	0.73	0.91
2010	2	3358	0.70	0.13	0.22	0.73	0.92
2015	2	4325	0.70	0.13	0.11	0.73	0.91
2020	2	2468	0.70	0.13	0.23	0.73	0.93
2025	2	1820	0.70	0.13	0.18	0.73	0.93
1985	3b	372	0.73	0.12	0.21	0.76	0.92
1990	3b	895	0.70	0.13	0.25	0.73	0.93
1995	3b	1023	0.71	0.13	0.21	0.74	0.91
2000	3b	2179	0.70	0.13	0.17	0.72	0.92
2005	3b	3104	0.70	0.13	0.19	0.72	0.92
2010	3b	3358	0.69	0.13	0.18	0.72	0.92
2015	3b	4325	0.69	0.13	0.08	0.72	0.93
2020	3b	2468	0.69	0.13	0.21	0.71	0.92
2025	3b	1820	0.69	0.13	0.10	0.72	0.93

Table 12: Cosine similarity summary statistics for natural sentences to prototype centroids by each sense and five-year interval, with heatmap shading to illustrate distributional stability.

Note. Year denotes the starting year of each five-year interval in the corpus period (1985-2025).

E Annotation Diagnostics and Threshold Estimation

This appendix describes the annotation diagnostics and base-rate-calibrated thresholding procedure used to determine cosine similarity cutoffs for sense identification. Threshold estimation proceeded in two stages. In Round 1, we assessed how gold-labeled senses were distributed across the cosine similarity space and evaluated whether a fixed purity-based decile threshold was viable. In Round 2, we enriched high-similarity cases for minority senses to enable reliable base-rate calibration of sense thresholds.

E.1 Human Annotation Protocol

Two expert annotators (NB and NH) independently judged whether each sampled sentence expressed the target sense of *schizophrenia*, using binary judgments (1 = expresses the sense; 0 = does not) based on Oxford English Dictionary definitions. A conservative gold label was assigned only when both annotators agreed that the sentence expressed the target sense. Both annotators have expertise in psychological science. Prior to annotation, they discussed sense definitions and agreed on conditions for assigning contextual usages to senses. Inter-annotator agreement in Round 1 was high (overall agreement: 99%; Cohen’s $\kappa = 0.98$), with only 6 disagreements out of 600 sentences (Sense 1: 99.5%; Sense 2: 99.5%; Sense 3b: 98.0%). Round 2 resolved all 3 disagreements (of 397), achieving 100% agreement. These disagreements were primarily due to the rarity of sentences assigned to senses 2 and 3b, which required careful review to ensure accurate sense assignment.

Sense 1: Psychiatric. Sense 1 was treated as the default meaning in orthodox psychiatric and scientific understanding of a mental disorder. A sentence was labeled as Sense 1 if it referred to: (i) scientific research, (ii) medications, (iii) other legitimate mental disorders like bipolar or autism (iv) official subtypes (e.g., “paranoid schizophrenia”), (v) support groups (vi) diagnosis (people cannot get diagnosed with split personality). In short, if the context was a person or people with a mental health problem annotators assumed Sense 1 unless the sentence was (a) clearly referencing something like split personality, or (b) it was very ambiguous and might well refer to the split personality sense. In line with OED definitions and prevailing psychiatric usage, Sense 1 was treated as the default interpretation in psychiatric and medical contexts unless there was clear evidence for an alternative sense.

Sense 2: Split-Personality. Sense 2 corresponds to the lay misconception equating schizophrenia with “split personality.” Sentences invoking this notion were labeled as Sense 2, including cases where the misconception was explicitly negated (e.g., “*Schizophrenia is not being two different people.*”(id: 2182bcb) or “*It’s not “split personality” Joanne Barreno is the mother of two adult children with schizophrenia, and for many years was a leading mental health consumer advocate locally.*”(id: 5b0a57cc). These sentences were accepted on the grounds that the new incorrect sense is distributionally invoked.

Sense 3b: Metaphorical. Sense 3b captured metaphorical uses of *schizophrenia* to denote inconsistency, contradiction, or incoherence in abstract systems (e.g., politics, markets, institutions).

E.2 Annotation Diagnostics and Threshold Estimation

Threshold estimation proceeded in two stages: Round 1 and Round 2. Round 1 diagnosed the distribution of gold-labeled senses across the cosine similarity space, evaluating whether a fixed purity-based decile rule could be applied. Round 2 enriched high-similarity cases for minority senses, enabling reliable base-rate calibration.

E.2.1 Round 1: Stratified annotation diagnostics

Round 1 sampled 20 sentences from each cosine similarity decile (1–10) across all years (1985–2025) for each sense. This stratified design ensured coverage from minimally to maximally prototypical usages across the ranked similarity space. The decile-based purity rule (80% sense purity per decile) was satisfied only for Sense 1 (psychiatric). No decile met this criterion for the minority senses (Senses 2 and 3b), reflecting their low prevalence in the corpus. Table 13 summarizes the distribution of gold-labeled instances across senses.

Sense	Gold examples	Average decile	Average cosine
Sense 1 (psychiatric)	180	5.5	0.693
Sense 2 (split-personality)	2	2.0	0.836
Sense 3b (metaphorical)	5	3.8	0.772

Table 13: Round 1 gold annotation summary across senses. Average decile and cosine similarity indicate where true positive instances occur in the ranked cosine space.

E.2.2 Round 2: Top-decile enrichment and base-rate calibration

Because the $\geq 80\%$ purity criterion could not be satisfied for the minority senses in Round 1, we conducted a second annotation round designed to better characterize high-similarity regions for Sense 2 (split-personality) and Sense 3b (metaphorical). Specifically, we focused on the top cosine-similarity decile (90th–100th percentiles) for each sense. This decile was further stratified into ten 1-percentile bands, from which up to 20 sentences per band were randomly sampled, yielding 200 annotated sentences per sense.

Using annotations from both rounds, we estimated the base rate of each sense in the full corpus using a two-stage stratified estimator following:

$$\hat{p}_s = 0.9 \cdot \hat{p}_{s,\text{bottom9}} + 0.1 \cdot \hat{p}_{s,\text{top}},$$

where $\hat{p}_{s,\text{bottom9}}$ denotes the observed proportion of the sense in the bottom nine cosine-similarity deciles estimated from Round 1 ($n = 180$), and $\hat{p}_{s,\text{top}}$ denotes the observed proportion of the sense in the top decile estimated from Round 2 ($n = 200$), scaled to represent the full decile mass.

For Sense 1 (psychiatric), no additional Round 2 enrichment was required because high-purity regions were already identified in Round 1; accordingly, only Round 1 estimates were used for its calibration. Finally, cosine similarity thresholds were selected by matching each sense’s estimated base rate to the ranked similarity distribution. For each sense, the threshold was defined as the cosine value at which the number of retained corpus sentences equaled $\hat{p}_s \times N$, where $N = 19,544$ is the number of sentences in the scored candidate set that entered the sense-scoring stage of the pipeline. Table 14 reports the final thresholds for each sense.

Sense	\hat{p}	Target k	Threshold cosine	Method
Sense 1 (psychiatric)	0.800	15,635	0.588	Base-rate calibrated (Round 1)
Sense 2 (split-personality)	0.012	235	0.884	Base-rate calibrated (Round 2)
Sense 3b (metaphorical)	0.031	606	0.873	Base-rate calibrated (Round 2)

Table 14: Cosine similarity thresholds selected to match estimated sense base rates in the full scoring set. Thresholds correspond to the k -th highest cosine score, where $k = \hat{p} \times N$.

Note. Because thresholds are calibrated to match estimated prevalence rather than to maximize recall, some gold-labeled instances may fall below threshold by design.

F Sense Proportions Diagnostics

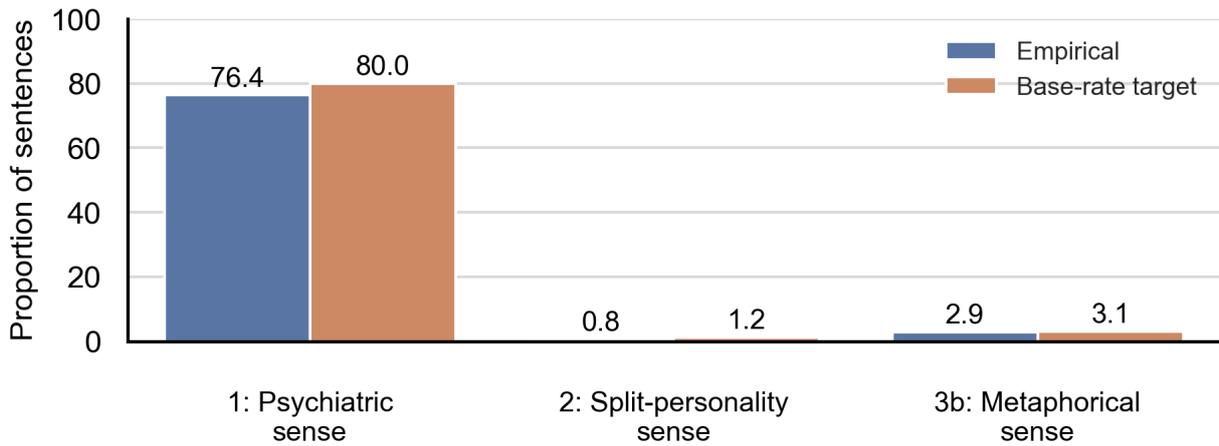


Figure 6: Global proportions versus base-rate targets.

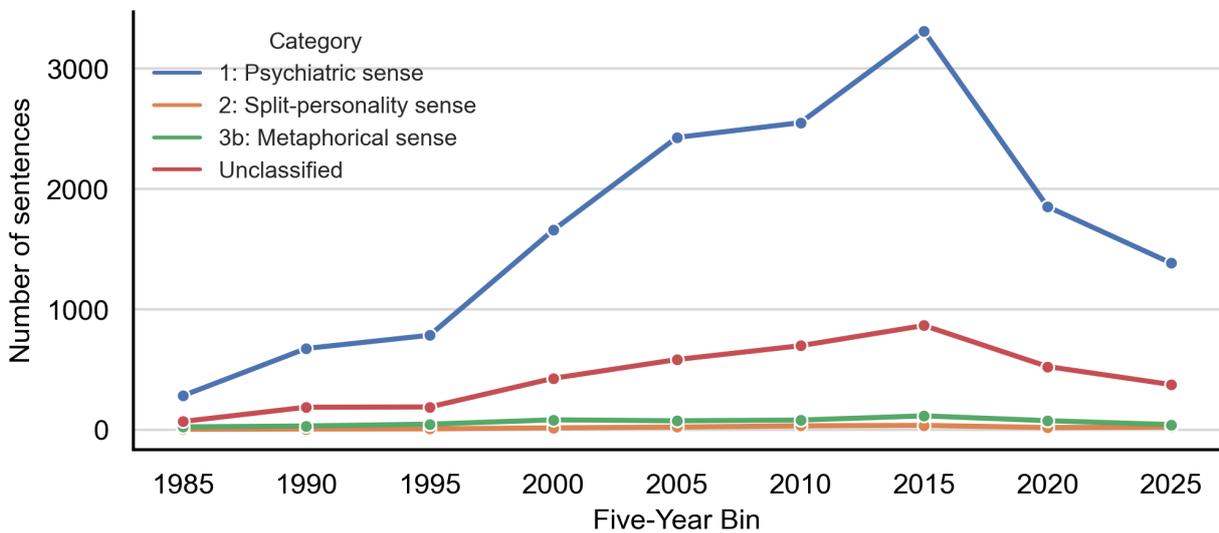


Figure 7: Sense counts over time (assigned senses and unclassified).

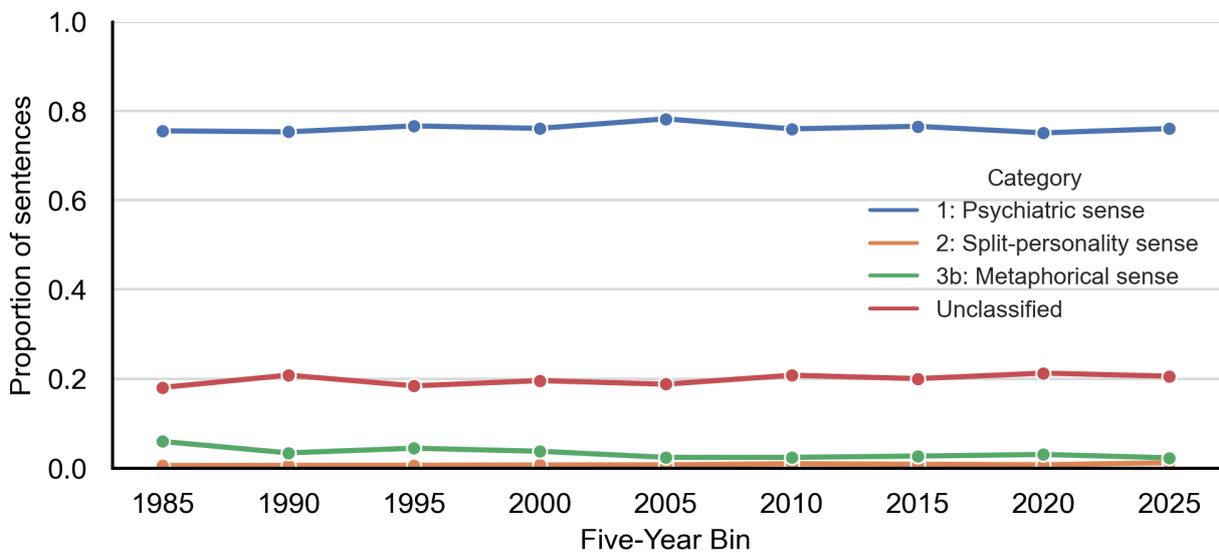


Figure 8: Sense proportions over time (base-rate-calibrated, including unclassified).

G Threshold exemplars for interpreting sense proportions

To help interpret the sense-prevalence plots, we present example sentences drawn from the same scored corpus used to estimate sense proportions. Sentences were assigned to senses using the same cosine-similarity thresholds applied in the quantitative analysis: for each sense, a sentence was considered an instance of that sense only if its similarity score exceeded the calibrated threshold, and each sentence was assigned to at most one sense based on its strongest match. We report two types of examples for each sense. First, borderline-positive sentences lie just above the threshold and illustrate the kinds of usages that are minimally included in a sense category. Second, high-confidence sentences have much higher similarity scores and serve as clear, representative examples of each sense. Tables 15 and 16 show these examples for Senses 1, 2, and 3b.

Sense	Bin	cos	τ	Δ	Sentence
Sense 1	2000–2004	0.59	0.59	0.000000	The project, based on Sylvia Nassar’s book, tells the true story of John Forbes Nash Jr., a mathematical genius with matinee-idol looks who suffered from schizophrenia but miraculously recovered and later received a Nobel Prize.
Sense 2	2015–2019	0.88	0.88	0.000	Perhaps not surprisingly, those who say they or someone else in their household faces an emotional or mental disability are especially likely to believe autism (75%), schizophrenia (67%) and depression (57%) should be considered qualifying conditions.
Sense 3b	2010–2014	0.87	0.87	0.00002	But WellPoint made no change to its coverage policies after that study, in part because the study was only for patients with schizophrenia, whereas the drugs are also commonly used in patients with bipolar disorder and depression.

Table 15: Borderline-positive exemplar sentences for each sense of *schizophrenia*. For each sense, sentences were scored by cosine similarity to the corresponding sense prototype and retained only if they exceeded the calibrated similarity threshold (τ). The examples shown are the first sentences lying just above this threshold (i.e., with the smallest positive margin, $\Delta = \text{cos} - \tau$), illustrating the types of usages that are minimally included under the classification criterion.

Sense	Bin	cos	τ	Δ	Sentence
Sense 1	2025–	0.92	0.59	0.33	Another defendant, M.H., a 65-year-old with schizophrenia, diabetes and severe asthma, was charged with misdemeanor crimes in April following a confrontation with a neighbor, the complaint said.
Sense 2	2020–2024	0.92	0.88	0.031	Inslee likened Trump’s response to “schizophrenia.”
Sense 3b	2020–2024	0.93	0.87	0.057	"The idea that schizophrenia means a split mind has contributed to a widespread belief in more formal contexts that the condition is like dissociative identity disorder, previously called split or multiple personality disorder, said Dr. Daniel Weinberger, dire. . ."

Table 16: Top-confidence exemplar sentences for each sense of *schizophrenia*. For each sense, sentences were scored by cosine similarity to the corresponding sense prototype and retained only if they exceeded the calibrated similarity threshold (τ). The examples shown are those with the largest margin above threshold ($\Delta = \text{cos} - \tau$), and therefore represent the clearest, highest-confidence instances of each sense under the classification procedure.

H Intra-sense heterogeneity within the psychiatric sense

To assess whether increases in semantic Breadth could arise from variation *within* the dominant psychiatric sense, we conducted an exploratory clustering analysis over Sense 1-assigned usages. Sentence embeddings were projected using UMAP and clustered with HDBSCAN to avoid k -selection. This analysis reveals several recurring subclusters corresponding to common psychiatric discourse contexts (e.g., diagnosis, treatment, institutional care), illustrated by representative high-margin exemplars in Table 17. While the overall semantic space remains temporally overlapping across decades (Fig. 10), distinct regions of the space are evident when colored by cluster assignment (Fig. 11). The relative prevalence of major subclusters demonstrates modest reweighting across five-year bins (Fig. 9), reflecting shifts in the relative frequency with which different psychiatric discourse contexts are invoked over time. Importantly, these subclusters are not treated as diagnostic subtypes (e.g., paranoid schizophrenia), but rather as recurring psychiatric discourse contexts whose relative salience changes over time within a stable semantic sense.

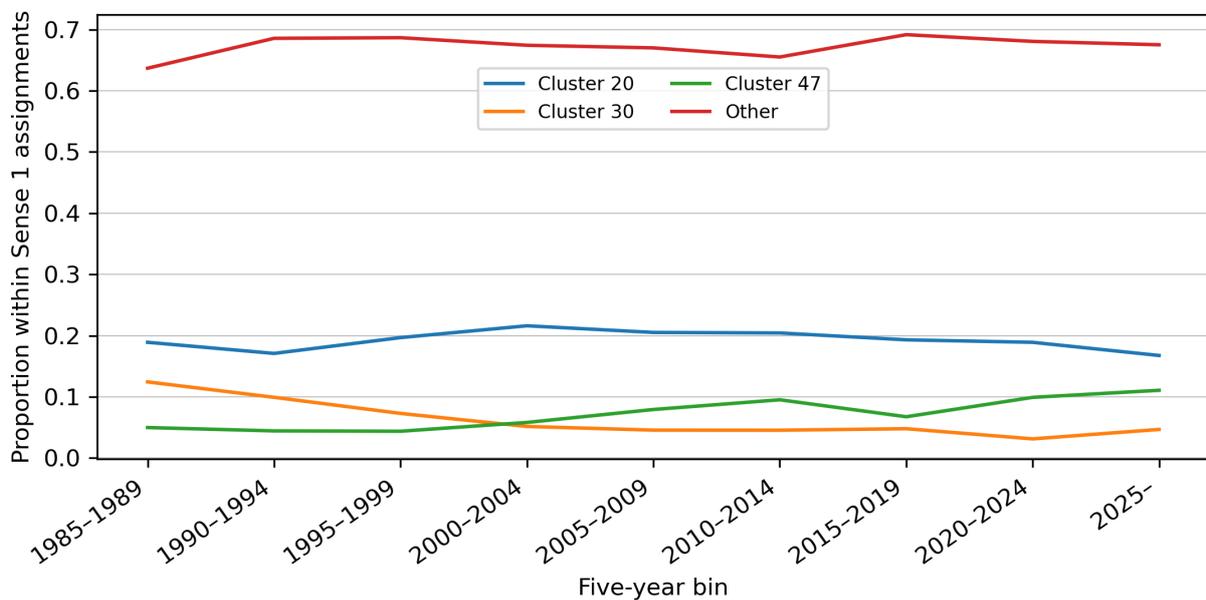


Figure 9: Relative prevalence of the three largest HDBSCAN subclusters within Sense 1 across five-year bins. Proportions are normalized within Sense 1 assignments per bin; remaining clusters are grouped as *Other*.

Cluster	Year	cos	Δ	Sentence
20	1985	0.86	0.28	All three Boston residents, each of whom had histories of schizophrenia, died while in seclusion rooms at the center.
30	1990	0.89	0.30	Schizophrenia, Paulus told the jury today, was one of the diagnoses given to S. as she moved from treatment to treatment. . .
12	2005	0.75	0.16	Through the end of last year, there were 15 potential new drugs for schizophrenia in human psychiatric testing. . .
47	1995	0.86	0.27	Perhaps because the symptoms can often be frightening to watch... schizophrenia remains largely misunderstood by the general population.
7	2010	0.81	0.22	Born Bernard Schwartz, Curtis was the Hungarian-Jewish son of a tailor father and a mother later diagnosed with schizophrenia.

Table 17: Representative high-margin exemplar sentences illustrating internal substructure within Sense 1 assignments. Exemplars are selected automatically and shown for interpretive illustration only; subclusters are not treated as distinct senses.

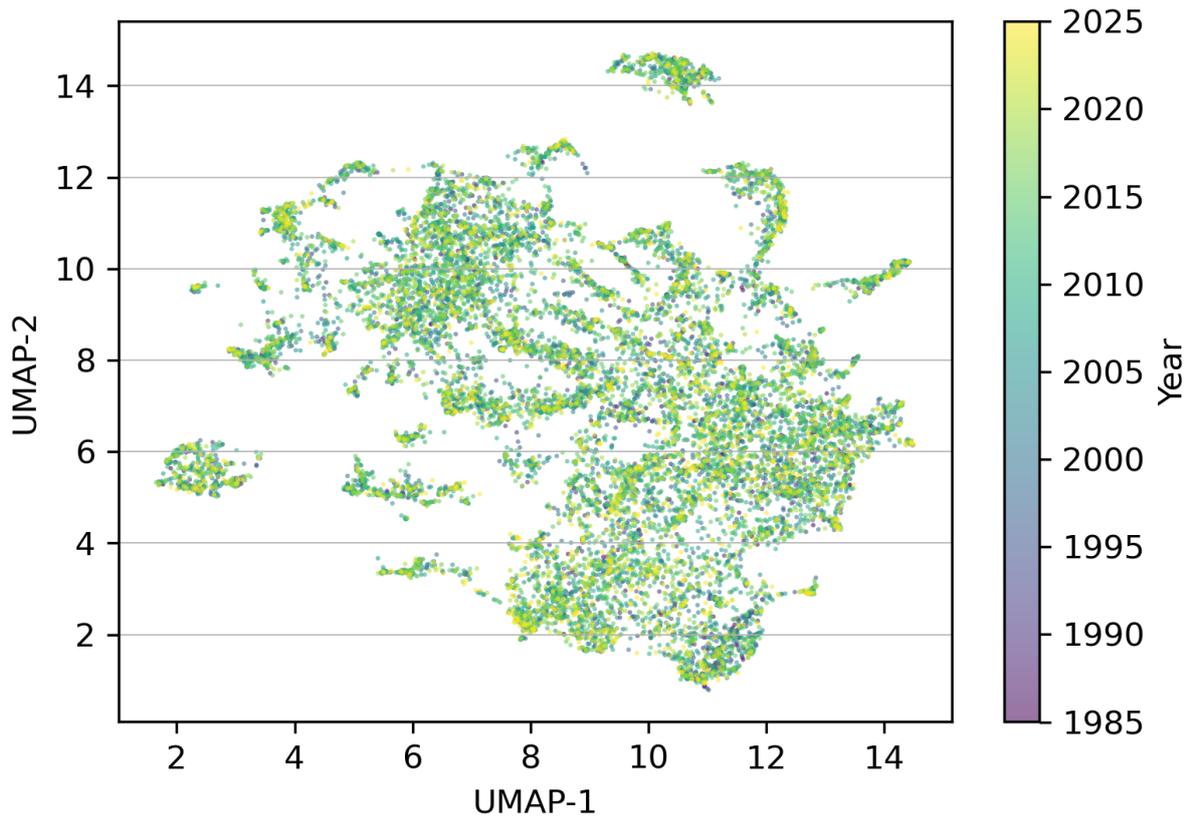


Figure 10: UMAP projection of Sense 1-assigned usages colored by year. Points from different decades largely overlap, indicating temporal continuity within the psychiatric sense.

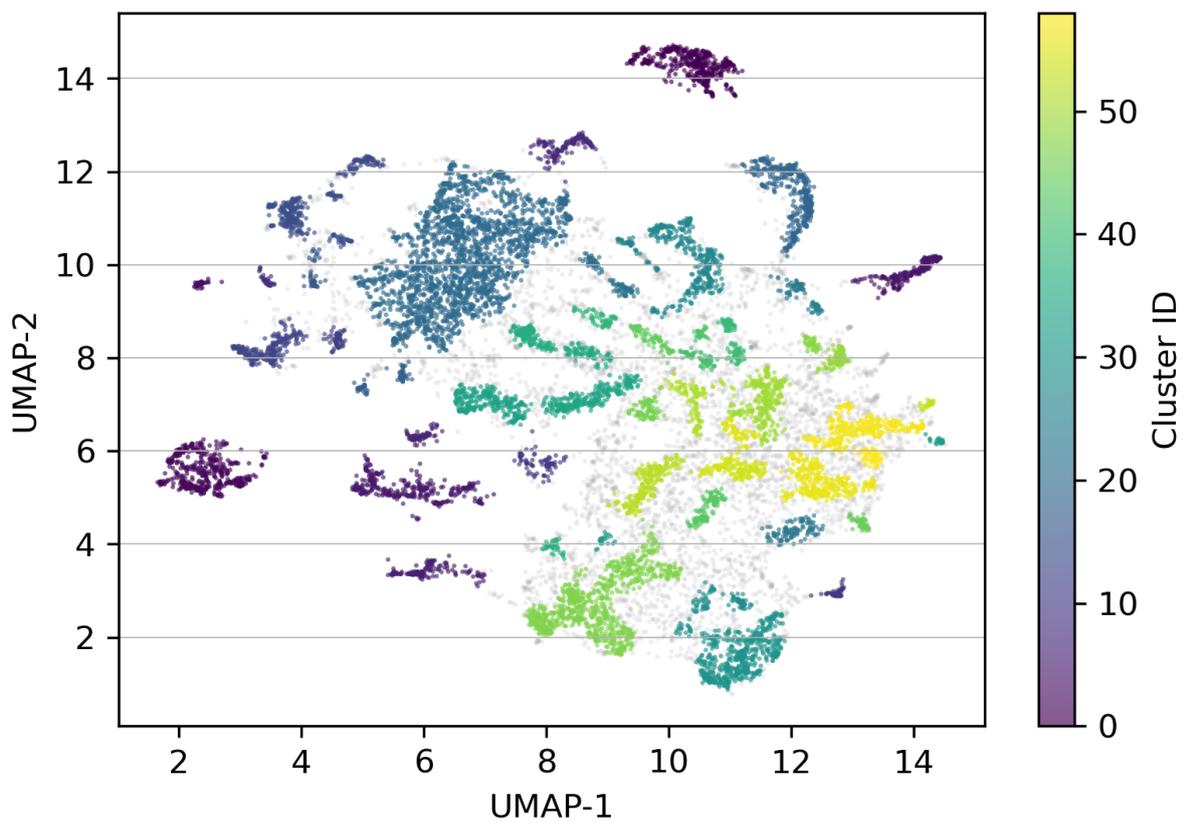


Figure 11: UMAP projection of Sense 1-assigned usages colored by HDBSCAN cluster. Gray points indicate noise assignments.

Using Correspondence Patterns to Identify Irregular Words in Cognate sets Through Leave-One-Out Validation

Frederic Blum^{1,2} Johann-Mattis List^{1,2}

¹Department of Linguistic and Cultural Evolution,

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

²Chair for Multilingual Computerlinguistics, University of Passau, Passau, Germany

Correspondence: frederic_blum@eva.mpg.de

Abstract

Regular sound correspondences constitute the principal evidence in historical language comparison. Despite the heuristic focus on regularity, it is often more an intuitive judgement than a quantified evaluation, and irregularity is more common than expected from the Neogrammarian model. Given the recent progress of computational methods in historical linguistics and the increased availability of standardized lexical data, we are now able to improve our workflows and provide such a quantitative evaluation. Here, we present the balanced average recurrence of correspondence patterns as a new measure of regularity. We also present a new computational method that uses this measure to identify cognate sets that lack regularity with respect to their correspondence patterns. We validate the method through two experiments, using simulated and real data. In the experiments, we employ leave-one-out validation to measure the regularity of cognate sets in which one word form has been replaced by an irregular one, checking how well our method identifies the forms causing the irregularity. Our method achieves an overall accuracy of 85% with the datasets based on real data. We also show the benefits of working with subsamples of large datasets and how increasing irregularity in the data influences our results. Reflecting on the broader potential of our new regularity measure and the irregular cognate identification method based on it, we conclude that they could play an important role in improving the quality of existing and future datasets in computer-assisted language comparison.

1 Introduction

Recurring correspondence patterns are the cornerstone of the comparative method and historical linguistics (Anttila, 1972; Fox, 1995; Durie and Ross, 1996). They are built upon the definition of regular sound change as a process that occurs across the whole lexicon in a specified set of contexts (Osthoff

and Brugmann, 1878, XIII). However, regularity is often more an intuitive judgement than a quantified evaluation, and irregularity is more common than expected from the Neogrammarian model of language change (Ross and Durie, 1996; Nichols, 1996; Blust, 2022).

Previous computational approaches have used instances of regular sound change to infer the correspondence patterns of a comparative wordlist (List, 2019) or to compare the regularity of datasets before and after applying methods that alter the correspondence patterns in the data (Blum and List, 2023). Implicitly, they also form the basis for automated approaches of reconstructing proto-languages (List et al., 2023). Yet, the full utility that regularity provides is rarely exploited.

Here, we present a new measure that evaluates the regularity of correspondence patterns in a standardized comparative wordlist. By computing the balanced average recurrence of correspondence patterns, we establish a score that can be used to identify cognate sets without such recurrent patterns. Through different normalizations, we can also compare this measure across datasets. We illustrate how this workflow can be used with a new method that detects erroneous cognate judgements in comparative wordlists. Our approach measures the regularity of cognate sets and reports on the cases with a low score. We can then implement a leave-one-out validation and iterate through each cognate set to identify those word forms whose deletion improves the regularity of the cognate set.

2 Background

Sound change has been recognized as an overwhelmingly regular process (Osthoff and Brugmann, 1878). If a sound change occurs in the lexicon of a given language at a certain point in time, it usually affects all of the words that occur in a particular phonetic environment, leaving almost no

exceptions (Campbell, 1996). Based on this regularity, scholars have developed methods for the reconstruction of proto-languages which are not attested in written sources, but that can be inferred through the systematic comparison of related languages. Those reconstructed languages can then be used for further comparisons to establish genealogical links between different language families.

Regularity of sound change is the core principle of linguistic reconstruction. Linguists identify recurring correspondence patterns through the annotation and alignment of cognate sets, that is, sets of related words in a language family (Anttila, 1972; Fox, 1995). Those recurring correspondence patterns build the basis for the reconstruction of proto-phonemes. Due to the regularity of those patterns, they can also be used for predicting reflexes in languages where we have no explicit evidence of the lexical form. Such approaches have been used for targeted fieldwork (Bodt and List, 2021) or testing hypotheses of genealogical relationships between language families (Blum et al., 2024a). The regularity of correspondence patterns also provides the arguments to distinguish chance similarity or borrowings from true genealogical descent.

The regularity of sound change doesn't describe dogmatic laws, but rather the definition of a *specific type* of sound change that differs from irregular forms of change such as analogy and borrowings (Hoenigswald, 1978; Labov, 1981). Those are considered to be of minor importance to the overall regularity, which keeps forming the basis for linguistic reconstructions (Ross and Durie, 1996; Campbell, 1996). The cognate annotation for related languages based on the regularity principle is still considered the state-of-the-art for historical linguistics. Its applications in the 21st century include the manual reconstruction of proto-phonemes (Zariquiey et al., 2026), the relative dating of sound changes (Fries and Korobzow, 2024), the prediction of cognate reflexes (Blum et al., 2024a), and the preparation of comparative wordlists for various computer-assisted methods (Wu et al., 2020; Blum et al., 2024b).

Within computer-assisted workflows, comparative wordlists annotated for cognate sets are also used in phylogenetic studies that infer the temporal dimension of the diversification of language families, as well as their internal classification (Greenhill et al., 2020). Comparative wordlists can be built in a standardized way using the Cross-Linguistic Data Formats (CLDF, Forkel et al.

(2018)). The standardization includes linking languages to Glottolog (Hammarström et al., 2024) and phonetic segments to the Cross-Linguistic Transcription Systems (List et al., 2024). The standards established for LexiBank datasets (List et al., 2022; Blum et al., 2025a) also allow for a transparent annotation of cognate sets. For example, such cognate-annotated wordlists have been used to study the history of Austronesian (King et al., 2024), Bantu (Grollemund et al., 2015), Indo-European (Heggarty et al., 2023), Mixtecan (Auderset et al., 2023), Sino-Tibetan (Sagart et al., 2019), Uto-Aztecan (Greenhill et al., 2023), and other language families. However, no computational methods exist to evaluate the presented cognate coding, and more often than not, the data is taken as granted by reviewers and readers. This is problematic in many ways, since different interpretations of the same data often co-exist (Anderson et al., 2025; Kassian and Starostin, 2025), and the data should be considered a fundamental part of any statistical model (McElreath, 2020). However, it has not really been defined how the regularity can be measured, and how individual cognate sets can be evaluated with computer-assisted methods.

3 Materials and Methods

3.1 Cognate-Coded Comparative Wordlists

For our study, we use 20 datasets from LexiBank v2.1 (Blum et al., 2025a). The datasets are available in the Cross-Linguistic Data Formats (Forkel et al., 2018), and are mapped to Glottolog (Hammarström et al., 2024), Concepticon (List et al., 2025), and CLTS (List et al., 2024). The sample is part of the CogCore subset of LexiBank, which only includes datasets with manually annotated cognacy. We analyse the majority language family of each dataset and preprocess the data via LexiBench (Häuser et al., 2025), which provides a number of modifiable thresholds. We have only included datasets with at least ten languages from the same family and a minimum number of 140 concepts mapped to Concepticon (List et al., 2025). We present the resulting 576 languages from 19 language families in Figure 1, and the number of languages and cognate sets per dataset in Table 1.

One problem for the development of methods that exploit the regularity of correspondence patterns is the lack of datasets with manual alignments. The only datasets in our sample that include such manual alignments for the whole comparative

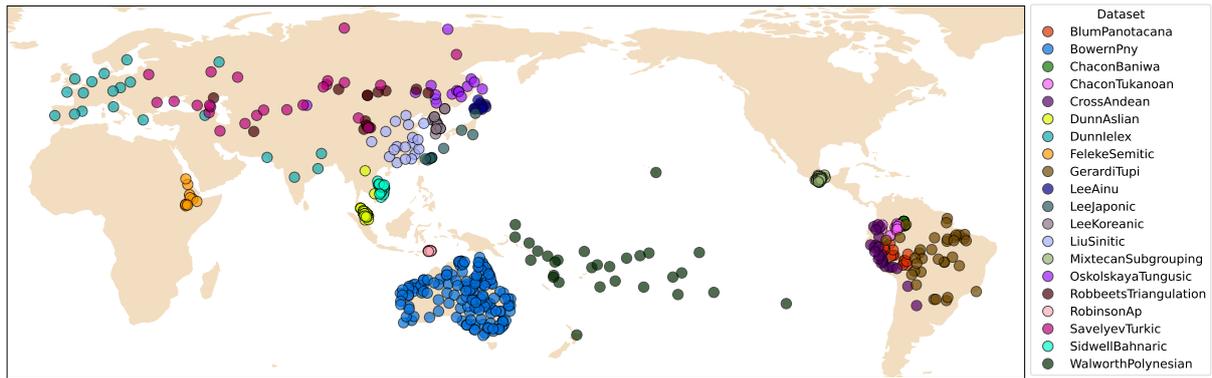


Figure 1: Map of all languages in the sample coloured per dataset.

wordlist are CrossAndean (Blum et al., 2023) and BlumPanotacana (Blum et al., 2024b). Tests on large benchmark datasets have shown that multiple phonetic alignment algorithms work rather well, achieving about 88% of identical columns on considerably divergent cognate sets (List, 2014). The benchmarks on which these algorithms were tested, however, did not reflect the kind of data that we observe in comparative wordlists, but rather consisted of hand-selected cognate sets from closely related dialect varieties. As a result, we expect that automated alignment methods serve as an additional source of irregularities in the data, since the data in comparative wordlists often comes from languages that not as closely related as the sampled dialect varieties on which the algorithms have been tested.

3.2 Measuring Regularity

Following the major findings made by linguists in the beginning of the 19th century, we know that sound change proceeds in an overwhelmingly regular manner (Osthoff and Brugmann, 1878; Hoenigswald, 1978). So far, however, quantitative measures for this particular regularity are lacking. A valid regularity measure for comparative wordlists must account for three domains, in which different notions of regularity play an important role. First, there are individual correspondence patterns. These are abstract representations of the columns (sites) of all alignments of the cognate sets in a given comparative wordlist. Based on their mutual compatibility, the individual alignment sites can be grouped together into correspondence patterns (List, 2019). The alignment sites that constitute a correspondence pattern can be counted directly, and we assume that regularity of individual patterns can be represented through frequency. Second, there are the cognate sets, which

are groups of words that were found to be related. The segments within each cognate set can be phonetically aligned, creating the alignment sites mentioned above. For each cognate set, we would like to measure how regular the combination of individual correspondence patterns underlying a particular alignment is compared to other alignments in the same dataset. Third, there is the dataset as a whole, which should display a certain balance, both in the number of regular cognate sets and the number of regular correspondence patterns.

The grouping of alignment sites into correspondence patterns is shown in Figure 2. Here, we see the cognates from four artificial languages for three different semantic concepts. The segments in all cognate sets are aligned phonetically. The individual columns (alignment sites) are then grouped into correspondence patterns based on their compatibility. For example, pattern I has the same reflex /k/ in all languages. In the case of concept B, the missing value for Language 4 is imputed based on the compatibility of the individual sites (List, 2019). Similarly to pattern I, pattern II features /n/ in all languages, with an imputed value in Language 4 for Concept B. Pattern III is distinguished from pattern I based on the contrastive reflex in Language 1, where we have /x/ (III) instead of /k/ (I).

For our analysis, we assume that cognates, alignments, and correspondence patterns have been inferred or computed in advance. Correspondence patterns are represented as a tuple of the length of the language varieties in a given wordlist. For each language, either the sound that would be expected in a regular alignment site is listed, or a specific symbol for missing data, in case the pattern cannot be fully resolved. Patterns are linked to individual alignment sites. In turn, the alignment sites can be linked to one or more patterns. Since the indi-

Dataset	Family	Source	Languages	Concepts	Cognates
BlumPanotacana	Pano-Tacanan	Blum et al. (2024b)	17	419	487
BowernPny	Pama-Nyungan	Bowern (2024)	175	238	2171
ChaconBaniwa	Arawakan	Chacon (2024a)	14	217	263
ChaconTukanoan	Tucanoan	Chacon (2024b)	16	141	141
CrossAndean	Quechuan	Blum et al. (2023)	34	150	245
DunnAslian	Austroasiatic	Dunn (2024a)	31	145	369
DunnLelex	Indo-European	Dunn (2024b)	20	205	485
FelekeSemitic	Afro-Asiatic	Feleke (2021)	21	150	273
GerardiTupi	Tupian	F. Gerardi and Reichert (2024)	37	243	395
LeeAinu	Ainu	Lee and Hasegawa (2025)	19	195	289
LeeJaponic	Japonic	Lee and Hasegawa (2024)	56	197	382
LeeKoreanic	Koreanic	Lee (2024)	14	206	230
LiuSinitic	Sino-Tibetan	Lilj et al. (2024)	19	202	306
MixtecanSubgrouping	Otomanguean	Auderset and Greenhill (2024)	84	224	436
OskolskayaTungusic	Tungusic	Oskolskaya et al. (2022)	21	254	498
RobbeetsTriangulation	Mongolic	Robbeets (2025)	15	253	415
RobinsonAp	Timor-Alor-Pantar	Robinson and Holton (2024)	13	217	256
SavelyevTurkic	Turkic	Savelyev and Robbeets (2024)	31	254	508
SidwellVahnaric	Austroasiatic	Sidwell (2024)	24	199	402
Walworthpolynesian	Austronesian	Walworth (2024)	31	210	511

Table 1: The datasets used for the experiment, including their number of languages, concepts, and cognate sets.

	Concept A					Concept B				Concept C				
	I	V	II	V	III	I	V	II	V	III	V	I	V	II
Language 1	k	a	n	o	x	k	e	n	a	x	o	k	e	n
Language 2	k	a	n	o	k	k	i	n	a	k	o	k	i	n
Language 3	k	a	n	o	k	k	i	n	a	k	o	k	i	n
Language 4	k	a	n	o	k	∅	∅	∅	∅	k	o	k	i	n

Figure 2: Artificial example for three cognate sets across four languages with their aligned segments. The columns are numbered and coloured according to their correspondence patterns. The vowel patterns are not further distinguished since all examples in the text refer to the patterns I-III.

vidual alignment sites may show varying amounts of missing reflex sounds in individual languages – given that cognate sets are rarely present in all languages of a family – they may show differential *compatibility* with alternative correspondence patterns (see List 2019 for details).

Our workflow for measuring the regularity of correspondence patterns in comparative wordlists consists of three steps. The starting point is an individual cognate set, represented as a phonetic alignment. This was illustrated in Figure 2. In the first step, we iterate over all *sites* in the alignment and identify all its compatible correspondence patterns. For each site, we select the pattern that is compatible with the highest number of individual alignment sites. This number reflects the *recurrence across sites* of the pattern. Calculating this recurrence for all sites of a given alignment yields a list consisting of the individual site recurrence scores, one for each column. To quantify the regu-

larity of a given cognate set, we first log-transform all individual recurrence scores and then compute the mean. This transformation helps to balance out potentially skewed distributions that may result in those cases where one pattern has a high recurrence score while all other patterns occur only once. If we take, for example, an alignment of four sites, with a site recurrence of [1 1 1 15], it would yield a higher absolute mean (4.5) than an alignment with a site recurrence of [2 2 2 2] (2). Taking the mean of the log-values will result in ranking the second set higher instead (0.67 vs 0.69). We can then take the exponential of this mean to arrive at a log-balanced average recurrence of individual sites within the cognate set (1.97 vs 1.99).

The third domain, regularity in the overall dataset, comes with an additional complication. Datasets differ in size with respect to the number of languages, concepts, and cognate sets they contain. In order to compare for regularity across datasets,

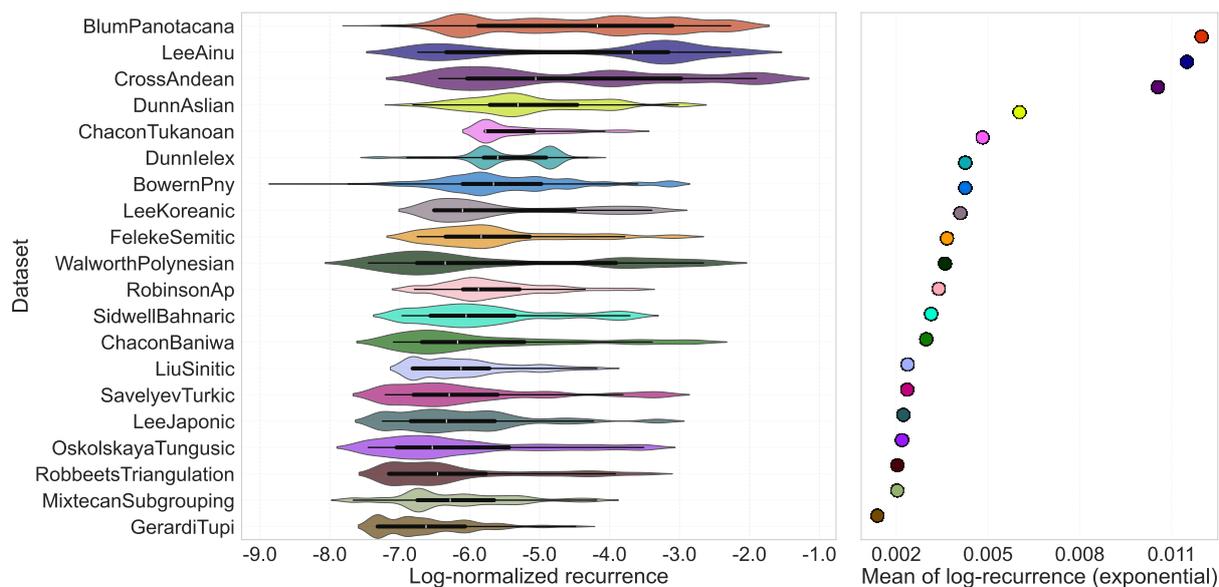


Figure 3: Overall regularity of the 20 datasets in the sample (y-axis) as indicated by measuring the average recurrence of sites in two ways. The left subplot shows the normalized and log-transformed recurrence of each site (x-axis). The right subplot shows the exponential of the mean of that log-transformed recurrence. This score can be interpreted as the balanced average pattern recurrence of an alignment site within each dataset.

we need to normalize the data. The first step here is to normalize the recurrence of each site through the total number of sites in the data. The normalized value is then log-transformed. In order to reach a regularity score that is comparable across datasets, we can take the exponential of the mean of the logarithm of the normalized values. This score describes the balanced average recurrence of a site within the dataset, and can be compared across datasets. The steps to reach this score can be summarized as follows: (1) counting the recurrence of alignment sites in the data, (2) normalize the count by the total number of sites, (3) log-transform the recurrence, and (4) take the exponential of mean of the log-recurrence (see Figure 3).

According to our new regularity measure, we can observe in Figure 3 that three datasets in our sample are far more regular than the others: CrossAndean, BlumPanotacana, and LeeAinu. What sets the first two apart qualitatively is that they are the only datasets for which we have manual alignments of the cognate sets. The fact that they score higher with respect to the recurrence of correspondence patterns is thus at least partially a display of the careful annotation of the data. Another important factor seems to be that in the first and the third dataset, the languages are very closely related. In those cases, irregularities are less likely to arise. With increasing distance between the languages, the irregularity in cognate sets tends to in-

crease, since there is more time for sporadic and irregular changes to occur.

The correspondence pattern measure suffers from another particular problem: with increasing language numbers, the methods will produce fewer and fewer regular correspondence patterns, since each language that is added to the selection may contain individual irregularities that then sum up to global irregularity. We will showcase this issue with simulated data. Another reason for the problem of apparently increasing irregularity is that correspondence patterns are not identical with proto-sounds, since individual regular sound changes may lead to a situation where one original proto-sound diversifies into several correspondence patterns. As a result, adding more languages to the comparison will almost always increase the number of correspondence patterns, due to individual sound changes that happen across branches.

3.3 Identifying Irregular Words

We confirm the utility of our measure through a new computational method that detects irregular words in cognate sets. For this purpose, we set up two experiments. In both experiments, we replace a single word form from an existing cognate set with a randomly generated irregular word. While the first experiment uses simulated regular data, the second experiment injects irregularities into real comparative wordlists (see Dessimoz et al.

2008 for this technique in the context of lateral gene transfer in biology and List 2015 for an application in the context of language contact in linguistics). In both experiments, we apply our new method to identify the most irregular form in a cognate set. In the experiment with the simulated data we show how pre-existing irregularity in correspondence patterns affects the accuracy of our method, and in the second we show how the method can successfully identify artificial forms that were injected into real cognate sets.

For the replacements, we first create a dictionary for each language that stores all its consonants and vowels. Then, we replace a randomly chosen word form in 20% of the cognate sets of each dataset. We replace each segment with a random entry from the dictionary for that language that is different to the original one, ensuring a form that has different segments than the original word form.

The basic workflow of our identification method is a leave-one-out validation. Within each of the cognate sets in the data where we have replaced a word form, we iterate through all its alignment sites and run the leave-one-out validation to compute the regularity. One word form at a time, we set the reflexes within the alignment to / \emptyset / and verify whether the altered alignment site can be matched to a compatible correspondence pattern in the data with higher recurrence than previously. By repeating this iteration for all alignment sites, we can identify the word form whose exclusion leads to the highest gain in regularity for that cognate set, measured as the average log-recurrence of its alignment sites. This identification directly relates to the second domain of regularity described above, which compares the average log-recurrence of sites between two cognate sets. We then verify whether the chosen form is the one which we have injected into the cognate set to evaluate the accuracy of our method. Both the replacement workflow and the evaluation are identical in both experiments.

3.3.1 Simulating Irregular Data

We use simulated data to show how pre-existing irregularity within comparative wordlists affects our method of identifying irregular word forms in cognate sets. For this purpose, we run code that simulates proto-forms for 200 concepts, based on a phoneme inventory including four vowels and ten consonants, and a CVCV syllable structure. For each of the ten simulated daughter languages, we simulate 0 to 2 regular phoneme mergers, each in-

Sample	Form	Sound Changes
A	r a l a	*/e/, */a/ > /a/
B	r a l e	
C	r e l e	*/e/, */a/ > /e/
D	r a k e	*/k/, */l/ > /k/
E	r a l e	
F	r a l e	
G	j i l e	*/i/, */a/ > /i/; */j/, */t/ > /j/
H	r a l i	*/e/, */l/ > /i/
I	w a l i	*/e/, */l/ > /i/; */w/, */t/ > /w/
J	r a l e	

Table 2: Phonetic alignments for word forms across ten simulated languages with random phoneme mergers for some languages.

volving the random selection of two phones from the vowel or consonant inventory. This ensures that we don't have any form identity in the cognate sets of the simulated data. The correspondence patterns are still fully regular, since we have the control over the data-generating process and apply all changes in all relevant instances. While neither the full regularity nor the sampled sound changes are a realistic scenario for real linguistic data, it allows us to verify the method in a controlled environment that involves an initial level of complexity. One such simulated cognate set is presented in Table 2.

For the experiment, we run ten different settings where we randomly replace between 5% and 50% of phones in the data. This leads to an artificial decrease in the overall regularity of the dataset through injecting random irregularities into the cognate sets, giving us a first indication how pre-existing irregularity in a comparative wordlist affects the inference of correspondence patterns.

3.3.2 Injecting Artificial into Real Data

The second experiment uses the 20 comparative wordlists presented in Table 1. All datasets include manually annotated cognate sets provided by the original study authors. The correspondence patterns are inferred using a subsample of languages from each dataset. We distinguish two settings: one that samples five languages, and one that samples ten. This step ensures comparability across datasets, which widely differ in the number of languages they feature. In order to compare the influence of this subsampling, we implement 100 runs for each dataset in each experimental condition and compare the distribution of the accuracy. In total, this results in 4,000 experimental trials.

This method has one big challenge: as we have already seen in Figure 3, the datasets are not per-

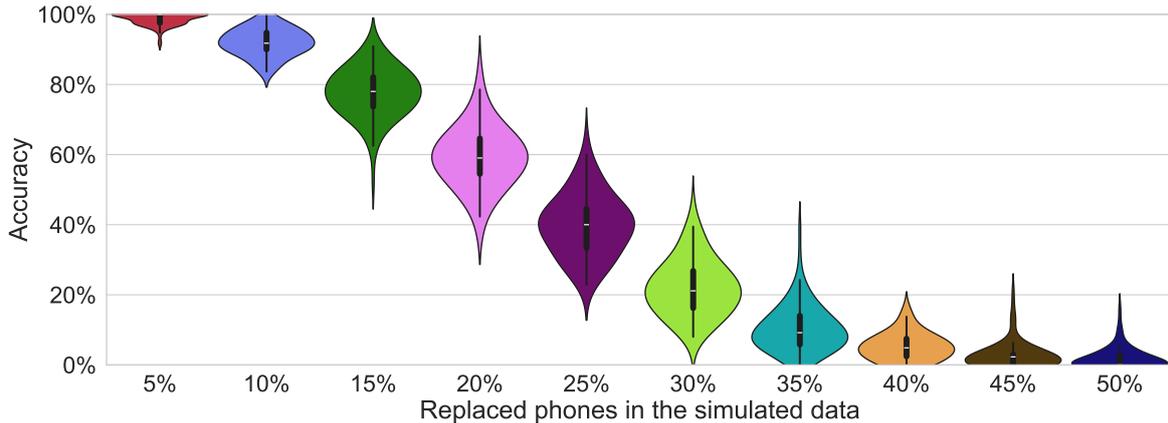


Figure 4: Accuracy for the leave-one-out validation with simulated regular data (y-axis) and random replacement of phones (x-axis) to simulate different levels of irregularity in comparative wordlists.

factly regular to begin with. The contrast with the simulated data shows that the real world of sound change is messy, and we do not expect our method to work perfectly. But once the task is defined, we can work on improving the methods and seek ways to counterbalance the factors contributing to the irregularity in the datasets. We will come back to this topic in the conclusions.

3.4 Implementation

We create Python code with a new dedicated package that provides the functionality for regularity computation. Cognate sets are already provided along with the individual cognate sets in our data sample. Where missing, phonetic alignments are computed with the help of the SCA algorithm (List, 2014), implemented in LingPy (List and Forkel, 2024). Correspondence patterns are inferred with the help of the CoPaR algorithm (List, 2019), implemented in LingRex (List and Forkel, 2022). All code and data needed to reproduce the results presented here are linked in the Data Availability Statement.

4 Results

4.1 Experiment with Simulated Data

In the first experiment we analyze the simulated data. This experiment not only shows how our method works in principle, but also how irregularity in the data affects its accuracy. The results for this experiment are presented in Figure 4.

The hypothesis that the accuracy of the method decreases with increasing irregularity in the data holds. Even for the otherwise perfectly regular data, accuracy drops to $\sim 90\%$ when replacing 10%

of phones. Starting at around 40% of replaced phones, our method drops below 10% accuracy on average. This indicates that the success of our method heavily depends on the initial regularity of the dataset. The less regular our dataset is in general, the more difficult it becomes to correctly identify the irregular forms. This simulation shows the complexity of the task once irregular correspondence patterns, expected to be abundant in the real data, creep in. But it also shows that our new measure can be applied successfully within computational methods that compare the regularity of competing alignment sites.

4.2 Replacing Sounds in Real Data

In the second experiment, we use 100 random subsample of five and ten languages for all 20 datasets. In each subsample, we replace one lexical word form in 20% of the cognate sets and verify whether our method correctly identifies the replaced form. The results for the experiment are presented in Figure 5. Our method reaches an overall accuracy of 85%, with the dataset means varying between 52% and 99%. This shows that the success of the method is largely dependent on the individual dataset. The results also show that the individual runs show a great variability in accuracy, indicating that the language sample and choice of cognate sets are important factors.

For all datasets, sampling only five languages instead of ten provides a considerable improvement in accuracy, confirming the potentially devastating impact of accumulating irregularities. The average accuracy for the five-language sample is 87.8% ($\sigma 20\%$), dropping to 79% ($\sigma 17\%$) for the 10-language sample. This confirms that the reg-

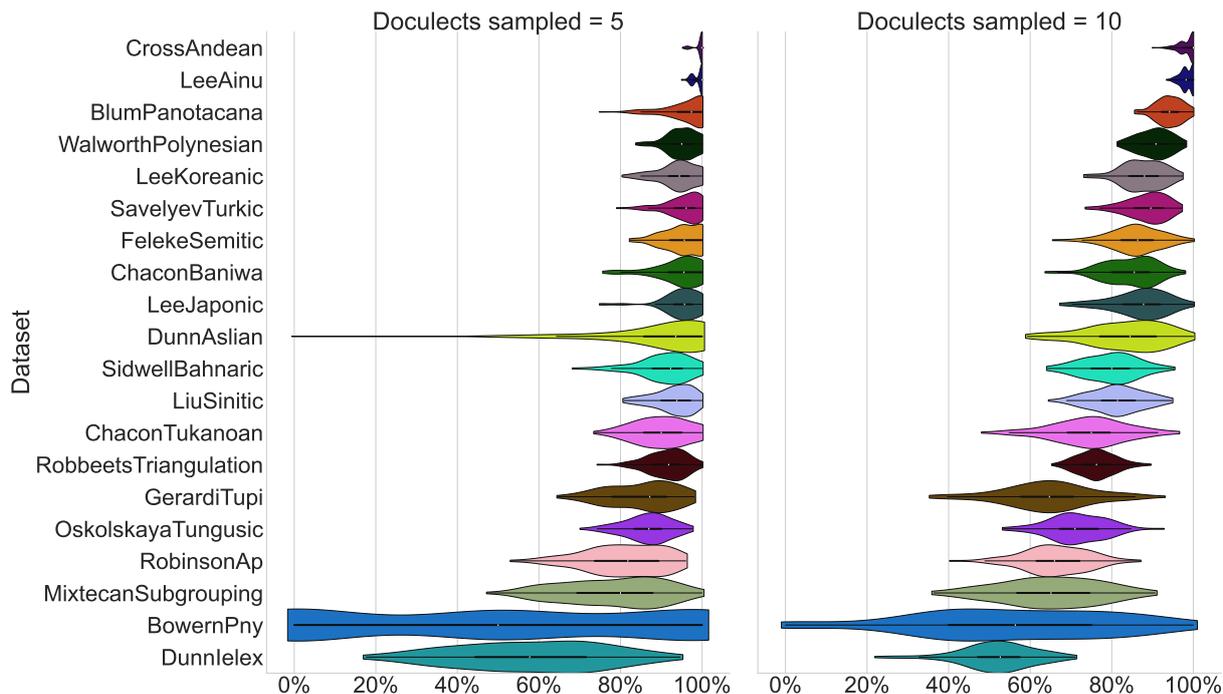


Figure 5: Distribution of experiment results for all datasets, split into the 5-doculects and 10-doculects sample settings. The boxplot presents the 50% distribution of all results.

ularity in the data is higher with fewer languages sampled, most likely because there are fewer individual reflexes disturbing the regularity of the correspondence patterns. The more languages we have in the data, the more possibilities for a single irregular site that disturbs the patterns arise. From this perspective, every additional language is another potential source for irregularities in alignment sites. We have seen a similar phenomenon with the simulated data. The more irregularity we find within individual cognate sets, the more difficult it is to identify the injected irregular form.

The same datasets top the accuracy that also showed the highest regularity in Figure 3 to begin with: *CrossAndean* (99%), *LeeAinu* (99%), and *BlumPanotacana* (95%). At the other end of the scale, there are two datasets that stand out with respect to the lowest amount of accuracy in the experiment: *BowermPny* (53%) and *Dunnlex* (55%). They are the only datasets with an overall average below 70%. Arguably, both include some of the most distantly related individual languages in our sample. *BowermPny* is also by far the largest dataset in the sample and represents one of the language families where automated approaches to language affiliation fail to achieve a large accuracy as well (Blum et al., 2025b). But the exact reasons why they perform worse cannot be answered here.

5 Conclusions

The regularity of language change is a complex phenomenon to which computer-assisted language comparison contributes a novel, quantitative perspective. We have presented a new measure that evaluates the average recurrence of correspondence patterns in standardized comparative wordlists. In contrast to previous implementations, this measure balances out several of the complexities of correspondence patterns – like their skewed distribution and the number of languages in the comparative wordlist – and is comparable across datasets. We have implemented this measure within a new computational method that successfully identifies irregular cognate sets and identifies those words that should potentially be excluded from the set. Two experiments using simulated and real datasets confirm the utility of the new measure and highlight different aspects of irregularity in comparative wordlists.

In the first experiment, using simulated data, our method reaches over 90% accuracy when randomly replacing up to 10% of the phones from the correspondence patterns. With increasing irregularities, the accuracy of our method decreases until dropping below 10% at 40% of replaced phones. The situation is different with the real datasets, where we find substantial differences between in-

dividual datasets. Those differences are influenced by the overall regularity in the dataset, as well as by the degree of relatedness between the languages in the data. On average, our method reaches 87.5% of accuracy when sampling five languages. The accuracy is lower when more languages are sampled.

There are many sources for irregularity, ranging from annotation errors, inconsistencies in the alignment algorithm, up to borrowings, or sporadic and irregular forms of sound change. Having identified the challenges, we can now focus at resolving them. First, we need to increase the data quality. While phonetic standardization through CLDF is a great first step, we need datasets that have also been manually aligned by experts. This will lead to an immediate increase in the regularity of the data, since no irregular patterns are introduced through automated alignment. Manual alignments could also lead to improved cognates, increasing in turn the regularity of alignment sites in the data, and resulting in better proto-language reconstructions.

We have also shown that sample size and language choice matter. The large deviations from the mean in the different experiment runs indicate that the accuracy heavily depends on the fact which languages and cognate sets have been sampled. The results also change with respect to how many languages are sampled, and the role of sample size in this kind of task will likely be a focal point of interest in the future. Through a more detailed analysis of this aspect, it should be possible to identify the individual languages which lead to an increased irregularity. These are likely the languages which represent key information for the history of the language family, since they might feature retentions and innovations not observed in other languages. By understanding which languages are especially complex in relation to the correspondence patterns can thus be of large benefit to historical linguists and provide important insights into the history of individual language families.

Our measure also provides a basis for additional applications. Regularity of cognate sets can and perhaps should be a commonly used tool for evaluating data in computer-assisted language comparison. For example, it could be used within computational methods that evaluate the cognate coding used for phylogenetic analysis and improve the cognate coding within such datasets. It could also point readers or reviewers to possibly problematic cognate judgements which require further explanation. It can also be used to help the traditional

reconstruction of language families by evaluating cognate judgements in the same way. We present such an application as part of our package, based on the leave-one-out comparison from the experiments. Another pathway can be the extension of the analysis to correspondence patterns involving proto-phonemes, and to identify those cognate sets that deviate from the proposed patterns.

Limitations

The main limitation is the availability of manually curated data for alignments, since the alignment of cognate sets is the crucial step previous to the inference of correspondence patterns. While there exist several datasets with manually annotated cognacy, there is only a handful of datasets with manual alignments. Another limitation is that the candidate selection of our method only works if there is exactly one entry causing the irregularity. While our method can always identify the irregular cognate sets, the leave-one-out methodology can only capture improvements if leaving out a single word forms leads to an increased regularity of a cognate set. If there are two irregular segments within the alignment, then no such improvement will be observed. In those cases, our method will be able to highlight the low regularity of the cognate set, but not be able to identify the word form(s) causing it.

Data Availability Statement

All code that has been used for this study is curated on Codeberg (<https://codeberg.org/calc/lingreg>, Version 0.1), containing a Python package and instructions how to run the code.

Funding

This research was supported by the Max Planck Society Research Grant CALC3 (JML, FB, <https://calclab.org>), the ERC Consolidator Grant *ProduSemy* (JML, Grant No. 101044282, see: <https://doi.org/10.3030/101044282>). All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher. The funders have/had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

- Cormac Anderson, Matthew Scarborough, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegen, Richard F. Strand, Geoffrey Haig, Ulrich Geupel, Martin Macak, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, Matthew Boutilier, Cassandra Freiberg, and 72 others. 2025. *The Indo-European cognate relationships dataset*. *Scientific Data*, 12(1541):1–27.
- Raimo Anttila. 1972. *An Introduction to Historical and Comparative Linguistics*. The Macmillan Company, New York.
- Sandra Auderset and Simon J. Greenhill. 2024. *CLDF dataset accompanying Auderset et al.’s “Subgrouping in a dialect continuum” from 2023*. Zenodo, Geneva.
- Sandra Auderset, Simon J. Greenhill, Christian T. DiCanio, and Eric W. Campbell. 2023. Subgrouping in a ‘dialect continuum’: A Bayesian phylogenetic analysis of the Mixtecan language family. *Journal of Language Evolution*, 8(1):33–63.
- Frederic Blum, Carlos Barrientos, Johannes Englisch, Robert Forkel, Simon J. Greenhill, Christoph Rzymiski, and Johann-Mattis List. 2025a. *Lexibank 2: pre-computed features for large-scale lexical data*. *Open Research Europe*, 5(126):1–27.
- Frederic Blum, Carlos Barrientos, Adriano Ingunza, and Johann-Mattis List. 2024a. Cognate reflex prediction as hypothesis test for a genealogical relation between the Panoan and Takanan language families. *Scientific Reports*, 14(30636):1–12.
- Frederic Blum, Carlos Barrientos, Adriano Ingunza, and Zoe Poirier. 2023. A phylolinguistic classification of the Quechua language family. *INDIANA - Anthropological Studies on Latin America and the Caribbean*, 40(1):29–54.
- Frederic Blum, Carlos Barrientos, Roberto Zariquiey, and Johann-Mattis List. 2024b. A comparative wordlist for investigating distant relations among languages in Lowland South America. *Scientific Data*, 11(92):1–9.
- Frederic Blum, Steffen Herbold, and Johann-Mattis List. 2025b. From isolates to families: Using neural networks for automated language affiliation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17915–17927, Vienna, Austria. Association for Computational Linguistics.
- Frederic Blum and Johann-Mattis List. 2023. Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–64, Dubrovnik, Croatia. Association for Computational Linguistics.
- Robert Blust. 2022. The Neogrammarian hypothesis and pandemic irregularity. *Journal of Historical Linguistics*, 12(2):167–193.
- Timotheus A. Bodt and Johann-Mattis List. 2021. Reflex prediction: A case study of Western Kho-Bwa. *Diachronica*, 39(1):1–38.
- Claire Bower. 2024. *CLDF dataset derived from Bower and Atkinson’s “Internal structure of Pama-Nyungan” from 2012*. Zenodo, Geneva.
- Lyle Campbell. 1996. On sound change and challenges to regularity. In Mark Durie and Malcolm Ross, editors, *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, pages 72–89. Oxford University Press, New York and Oxford.
- Thiago Chacon. 2024a. *CLDF dataset derived from Chacon et al.’s “Diversity of Arawakan languages” from 2019*. Zenodo, Geneva.
- Thiago Chacon. 2024b. *CLDF dataset derived from Chacon’s “A revised proposal of Proto-Tukanoan consonants and Tukanoan family classification” from 2014*. Zenodo, Geneva.
- Christophe Dessimoz, Daniel Margadant, and Gaston H. Gonnet. 2008. DLIGHT – Lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In Martin Vingron and Limsoon Wong, editors, *Research in Computational Molecular Biology*, pages 315–330. Springer, Berlin and Heidelberg.
- Michael Dunn. 2024a. *CLDF dataset derived from Dunn et al.’s “Aslian Languages” from 2013*. Zenodo, Geneva.
- Michael Dunn. 2024b. *CLDF Dataset derived from Dunn’s “IELEX” from 2012*. Zenodo, Geneva.
- Mark Durie and Malcolm Ross. 1996. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press, New York and Oxford.
- Fabrizio F. Gerardi and Stanislav. Reichert. 2024. *CLDF dataset derived from Gerardi and Reichert’s “The Tupi-Guaraní Language Family: A phylogenetic classification” from 2021*. Zenodo, Geneva.
- Tekabe Legesse Feleke. 2021. Ethiosemitic languages: Classifications and classification determinants. *Ampersand*, 8(100074):1–15.
- Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymiski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1):1–10.
- Anthony Fox. 1995. *Linguistic Reconstruction: An Introduction to Theory and Method*. Oxford University Press, Oxford.

- Simon Fries and Natalie Korobzow. 2024. [On the dating of sound changes and its implications for language relationship: The case of Proto-Yeniseian *p > Ket h-, Yugh f-](#). *Diachronica*, 41(4):525–555.
- Simon J. Greenhill, Hannah J. Haynie, Robert M. Ross, Angela M. Chira, Johann-Mattis List, Lyle Campbell, Carlos A. Botero, and Russell D. Gray. 2023. [A recent northern origin for the Uto-Aztecan family](#). *Language*, 99(1):81–107.
- Simon J. Greenhill, Paul Heggarty, and Russell D. Gray. 2020. [Bayesian phylolinguistics](#). In Richard D. Janda, Brian D. Joseph, and Barbara S. Vance, editors, *The Handbook of Historical Linguistics*, volume II, chapter 11, pages 226–253. Wiley, Hoboken.
- Rebecca Grollemund, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel. 2015. [Bantu expansion shows that habitat alters the route and pace of human dispersals](#). *Proceedings of the National Academy of Sciences*, 112(43):13296–13301.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. [Glottolog database \(v5.2.1\)](#). Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Luise Häuser, Robert Forkel, and Johann-Mattis List. 2025. [PyLexibench – generating data for Lexibench with a Python package](#). *Computer-Assisted Language Comparison in Practice*, 8(1):25–37.
- Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, and 14 others. 2023. [Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages](#). *Science*, 381(6656):eabg0818.
- Henry M. Hoenigswald. 1978. [The annus mirabilis 1876 and posterity](#). *Transactions of the Philological Society*, 76(1):17–35.
- Alexei S. Kassian and George Starostin. 2025. [Do ‘language trees with sampled ancestors’ really support a ‘hybrid model’ for the origin of Indo-European? Thoughts on the most recent attempt at yet another IE phylogeny](#). *Humanities and Social Sciences Communications*, 12(682):1–10.
- Benedict King, Simon J. Greenhill, Lawrence A. Reid, Malcolm Ross, Mary Walworth, and Russell D. Gray. 2024. [Bayesian phylogenetic analysis of Philippine languages supports a rapid migration of Malayo-Polynesian languages](#). *Scientific Reports*, 14(14967):1–12.
- William Labov. 1981. [Resolving the Neogrammarian controversy](#). *Language*, 57(2):267–308.
- Sean Lee. 2024. [CLDF dataset derived from Lee’s “Sketch of language history in the Korean peninsula” from 2015](#). Zenodo, Geneva.
- Sean Lee and Toshikazu Hasegawa. 2024. [CLDF dataset derived from Lee and Hasegawa’s “Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages” from 2011](#). Zenodo, Geneva.
- Sean Lee and Toshikazu Hasegawa. 2025. [CLDF dataset derived from Lee and Hasegawa’s “Evolution of the Ainu Language in space and time” from 2013](#). Zenodo, Geneva.
- Johann-Mattis List. 2014. [Sequence Comparison in Historical Linguistics](#). Düsseldorf University Press, Düsseldorf.
- Johann-Mattis List. 2015. [Network perspectives on Chinese dialect history](#). *Bulletin of Chinese Linguistics*, 8(1):42–67.
- Johann-Mattis List. 2019. [Automatic inference of sound correspondence patterns across multiple languages](#). *Computational Linguistics*, 45(1):137–161.
- Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, Christoph Rzymiski, and Robert Forkel. 2024. [CLTS. Cross-Linguistic Transcription Systems \(v2.3.0\)](#).
- Johann-Mattis List and Robert Forkel. 2022. [LingRex. Linguistic Reconstruction with LingPy \(v1.2.0\)](#).
- Johann-Mattis List and Robert Forkel. 2024. [LingPy. A Python library for quantitative tasks in historical linguistics. Version 2.6.13](#). University of Passau, Passau.
- Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymiski, Johannes Englisch, and Russell D. Gray. 2022. [Lexibank, a public repository of standardized wordlists with computed phonological and lexical features](#). *Scientific Data*, 9(1):1–16.
- Johann-Mattis List, Nathan W. Hill, Robert Forkel, and Frederic Blum. 2023. [Representing and computing uncertainty in phonological reconstruction](#). In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 22–32, Singapore. Association for Computational Linguistics.
- Johann-Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos, Christoph Rzymiski, Simon J. Greenhill, and Robert Forkel. 2025. [Concepticon v3.4.0. A Resource for the Linking of Concept Lists](#). Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Liú Lǐlǐ, Wáng Hóngzhōng, and Bǎi Yíng. 2024. [CLDF dataset derived from Liú et al.’s “Collection of basic words in Chinese dialects” from 2007](#). Zenodo, Geneva.

- Richard McElreath. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 2nd edition. Chapman and Hall/CRC, New York.
- Johanna Nichols. 1996. **The comparative method as heuristic**. In Mark Durie and Malcolm Ross, editors, *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, pages 39–71. Oxford University Press, New York and Oxford.
- Sofia Oskolskaya, Ezequiel Koile, and Martine Robbeets. 2022. **A Bayesian approach to the classification of Tungusic languages**. *Diachronica*, 1(39):1–31.
- Hermann Osthoff and Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*, volume 1. Hirzel, Leipzig.
- Martine Robbeets. 2025. *CLDF dataset derived from Robbeets et al.’s “Triangulation of Transeurasian languages” from 2021*. Zenodo, Geneva.
- Laura C. Robinson and Gary Holton. 2024. *CLDF dataset derived from Robinson and Holton’s “Internal classification of the Alor-Pantar language family” from 2012*. Zenodo, Geneva.
- Malcolm Ross and Mark Durie. 1996. **Introduction**. In Mark Durie and Malcolm Ross, editors, *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, pages 3–38. Oxford University Press, New York and Oxford.
- Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. **Dated language phylogenies shed light on the ancestry of Sino-Tibetan**. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.
- Alexander Savelyev and Martine Robbeets. 2024. *CLDF dataset derived from Savelyev and Robbeet’s “Internal Structure of the Turkic Language Family” from 2020*. Zenodo, Geneva.
- Paul Sidwell. 2024. *CLDF Dataset derived from the Bahnaric data in Sidwell’s “Austroasiatic dataset for phylogenetic analysis” from 2015*. Zenodo, Geneva.
- Mary Walworth. 2024. *CLDF dataset derived from Walworth’s “Polynesian segmented data” from 2018*. Zenodo, Geneva.
- Mei-Shin Wu, Nathanael E. Schweikhard, Timotheus A. Bodt, Nathan W. Hill, and Johann-Mattis List. 2020. **Computer-assisted language comparison: State of the art**. *Journal of Open Humanities Data*, 6(2):1–14.
- Roberto Zariquiey, Frederic Blum, and Pilar Valenzuela. 2026. **The forgotten mid-vowels: Reconstructing the history of the Proto-Pano vowel system**. *Knowledge Commons*, pages 1–35.

DHPLT: large-scale multilingual diachronic corpora and word representations for semantic change modelling

Mariia Fedorova¹, Andrey Kutuzov¹, Khonzoda Umarova²
mariiaf@ifi.uio.no, andreku@ifi.uio.no, ku47@cornell.edu
(equal contribution, authors sorted in alphabetical order)
¹University of Oslo (Norway), ²Cornell University (USA)

Abstract

In this resource paper, we present DHPLT, an open collection of diachronic corpora in 41 diverse languages. DHPLT is based on the web-crawled HPLT datasets; we use web crawl timestamps as the approximate signal of document creation time. The collection covers three time periods: 2011-2015, 2020-2021 and 2024-present (1 million documents per time period for each language). We additionally provide pre-computed word type and token embeddings and lexical substitutions for our chosen target words, while at the same time leaving it open for the other researchers to come up with their own target words using the same datasets.

DHPLT aims at filling in the current lack of multilingual diachronic corpora for semantic change modelling (beyond a dozen of high-resource languages). It opens the way for a variety of new experimental setups in this field.

1 Introduction

Computational data-driven diachronic semantic change modelling (tracing meaning shifts over time) naturally requires diachronic corpora: that is, texts annotated with their creation date. Once such datasets are obtained, one can compare the usage of target words (or any other linguistic phenomena) in different time periods, using any preferred change modelling method.

However, diachronic corpora of appropriate size and quality are not easy to find, especially permissively licensed. Most of the current lexical semantic change detection (LSCD) projects operate on the same small set of high-resource languages. For example, the seminal SemEval 2020 Task 1 on LSCD (Schlechtweg et al., 2020) was limited to English, German, Latin and Swedish. Later, LSCD benchmarks based on diachronic corpora for Italian (Basile et al., 2019), Greek (Perrone et al., 2019), Russian (Kutuzov and Pivovarova, 2021), Norwegian (Kutuzov et al., 2022a), Spanish

(Zamora-Reina et al., 2022), Chinese (Chen et al., 2023), Japanese (Ling et al., 2023), Finnish (Fedorova et al., 2024b), and Slovene (Pranjić et al., 2024) were presented, but not more than that (not for all languages the corpora themselves are publicly available). The field has to experiment with at most a dozen of languages, with Indo-European family strongly over-represented. This limits the scope of LSCD research, especially on *multilingual* semantic change effects.

To fill in this gap, we release **DHPLT** (‘Diachronic HPLT’): a set of standardized diachronic corpora for 41 languages of 12 different language families. Each language is represented with three time-dependent subsets, containing 1 million documents each. These documents are extracted from the web-crawled datasets by the HPLT project, specifically HPLT v3.0 (Oopen et al., 2025): thus, they are basically cleaned and filtered web pages in the target language. We use crawling timestamps as the signal for time period separation (see below).

In addition, we define a set of potentially interesting ‘target words’ for each language. For the DHPLT occurrences of these words, we produce a variety of semantic representations (static word2vec embeddings, token embeddings, lexical substitutions). This allows practitioners to start experimenting with multilingual LSCD immediately, without spending compute on re-creating these representations. At the same time, the availability of the original texts makes it possible to come up with other target word sets. **All the resources described in this paper are available at <https://data.hp1t-project.org/three/diachronic/>, sorted by language.**

2 Diachronic corpora out of HPLT

In prior work, diachronic resources for LSCD mostly were produced from existing historical corpora manually created by linguists: newspaper

archives, releases by national libraries, etc. Unfortunately, such resources are nearly non-existent for the majority of world’s languages: at least in anything resembling a standardized form. In an ideal world, fragmented national efforts in historical corpora creation could be unified and merged into a multilingual diachronic resource. But the amount of work required for such a project is well beyond the scope of this paper or any research group we are aware of. That’s why we instead suggest to rely on the Internet as the source of diachronic data.

World Wide Web contains hundreds of billions of documents in all existing languages and of varying quality (many documents consist of SEO keywords, machine-generated slop or price lists). At least two initiatives are currently crawling the WWW and saving representative slices of its state: Common Crawl¹ (CC) and Internet Archive² (IA). The HPLT project (Burchell et al., 2025) processes these web crawls by conducting language identification, deduplication, cleaning, etc, to produce language-specific corpora of competitive quality.³ Importantly, all its datasets are published under the Creative Commons CC0 license. HPLT v3.0⁴ is the specific data release we are using.

HPLT provides a lot of clean documents, but to create a diachronic corpus, we need to know the *date* when the document was created. It is impossible to label all the HPLT documents with the creation date manually. Sometimes, web pages do contain the date of their publication either in plain text or in some structured form. But this data is not reliable: it is perfectly possible for a web document to be published in 2024, but contain text created in 2001. Also, creating parsing rules for all sorts of HTML creation date labels would be an immense effort - with no guarantee that the result will fully reflect the diversity of the Web. Thus, we instead rely on a different time signal: web crawling time stamps. All the HPLT documents can be traced back to specific web crawls and they inherit the ‘timestamp’: that is, the exact date and time when a given web page was downloaded and saved.

Admittedly, these timestamps do not directly map to the creation date of the document: again, it is absolutely possible for CC or IA in 2024 to

download a web page created in 2001. But the timestamps do provide an ‘upper boundary’ of the creation date: if some text was crawled in 2015, there is no way for it to have been created later than 2015. Web crawl timestamps allow us to create diachronic datasets of a sort slightly different from ‘traditional’ diachronic corpora. Here, subsets for periods 1, 2 and 3 contain documents created *no later than 1, 2 or 3* respectively. Importantly, the subset 3 can still contain documents created in the earlier time periods: but not vice versa. For sure, this is less precise than manually labelled historical corpora: but we believe this still can be an important source of diachronic text data.

We aim at a diachronic dataset with more than two time periods, since this makes it possible to conduct research in long-term multipoint dynamics of semantic change (Kutuzov and Pivovarova, 2021). We also would like our time periods to be as comparable as possible in terms of the amount of data, and to be separated by at least some ‘gaps’, since this makes it easier to detect semantic change (Giulianelli et al., 2022). In order to choose the exact temporal spans, we analyse the distribution of documents in the HPLT v3.0 datasets by the year of crawling. Figure 1 in the Appendix shows these numbers for English and Georgian as an example. Our main observations are that 1) 2011 is the earliest crawl year, and the number of documents remains relatively low until 2017; 2) much more documents were crawled in 2020 and after, with peaks in 2020 and 2024 (the latest crawl year).⁵

Based on these observations, we come up with the following three time periods, each 2-4 years long, and with gaps of at least two years: 2011-2015 (*‘early time period’*), 2020-2021 (*‘Covid time period’*), 2024 (*‘most recent crawls’*). The three-time-period structure for the DHPLT is useful for studying and capturing linguistic innovation or the onset of semantic change at different points in time. With the crawl timestamps being the ‘upper bound’ on the document creation time, we can, for instance, observe the rise of the concept of ‘remote work’ in 2020-2021 and then look at its journey in 2024.

Note that our time bins are far from being the only possible choice. We consider them to be a sensible way of temporally splitting the existing HPLT data, but depending on the objective, other splits can make more sense. All the documents

¹<https://commoncrawl.org/>

²<https://archive.org/>

³Another project from which one could extract diachronic web corpora is FineWeb 2 (Penedo et al., 2025).

⁴<https://hplt-project.org/datasets/v3.0>

⁵Interestingly, 2013 is a rather rare crawl year in HPLT v3.0: many languages have no documents crawled in 2013.

in our datasets are accompanied with full timestamps, so anyone can produce their own subsets of DHPLT: for example, more fine-grained. It is also possible to use our open source code⁶ to reproduce DHPLT from the original HPLT data with any desired changes.

We produce three subsets of the HPLT v3.0 datasets containing documents crawled during the time periods above. But first, we need to choose what *languages* DHPLT will contain.

2.1 Language selection

The original HPLT v3.0 datasets feature 198 languages, which is way too many for our purposes. Our selection of languages for DHPLT is based on the following criteria:

1. Language must have at least 0.5 million documents in each of the time periods above: smaller languages do not provide sufficient amount of data and also are more error-prone with regards to language identification.
2. There should exist a corresponding HPLT v3.0 T5 monolingual encoder-decoder language model⁷ (Open et al., 2025): we use these models to generate token embeddings in 4.

As a result, we come up with a set of 41 languages. Table 1 in the Appendix lists them along with their ISO codes (augmented with the writing system code) and the corresponding number of documents in each of the three time periods.

2.2 Data extraction pipeline

For each of the languages, we construct three time-specific corpora by randomly sampling 1 million documents from the HPLT v3.0 dataset corresponding to the given language and time period. Where less than 1 million documents are available, we only sample 0.5 million. The resulting diachronic corpora are published as zstd-compressed JSONL files, following the HPLT format, with the total size ≈ 170 GB (for comparison, the full size of HPLT v3.0 is 50 TB), and ≈ 59 billion words.

These diachronic corpora can already be used for multilingual LSCD research. However, we also provide more ‘refined’ data for practitioners: namely, semantic representations (which can be computationally expensive for academic researchers to produce) for pre-defined sets of ‘target words’. They

⁶https://github.com/lrgoslo/scdisc_hplt

⁷<https://hf.co/collections/HPLT/hplt-30-t5-models>

are described in the next sections. One can think about them as an *example* of what sorts of experimental setups are possible with DHPLT.

3 Target word selection

For each language we select a subset of the vocabulary – target words – representations of which would be part of DHPLT. Our primary objective in selecting target words is to narrow down the full corpus vocabulary while keeping as many words that would be of interest to lexical semantic change researchers as possible.

Starting from the T5 model vocabulary corresponding to each given language, we filter out word pieces and infrequent tokens, leaving only words that appear as nouns, verbs or adjectives and are written in the language’s main script. Please refer to Appendix B for full details of our target word selection process.

Our selection pipeline yields a set of target words for each DHPLT language, with the average size of ≈ 18600 (HPLT T5 models’ vocabulary size is 32768). Figure 2 shows the distribution of target word counts across languages.

Target lemmas We additionally lemmatize each of the resulting target words. For instance, in English distinct target tokens ‘thread’, ‘Thread’, and ‘threads’ share one common lemma ‘**thread**’. We later use lemmas to merge word representations into more linguistically-informed groupings.

4 Target word representations

Once the language-specific target words are defined, we produce a number of different semantic representations for their occurrences in the DHPLT corpora. These representations can be directly used by LSCD practitioners to evaluate or train semantic change models on the three DHPLT time periods.

4.1 Contextualized word embeddings

Contextualized token embeddings are widely utilized in lexical semantic-change research (Periti and Tahmasebi, 2024; Umarova et al., 2025, *inter alia*). They can serve both as direct representations that are later averaged into prototypical embeddings (Periti and Montanelli, 2024) and as a basis for constructing clusters corresponding to different ‘sense nodules’ (Martinc et al., 2020; Kutuzov et al., 2022b). For our DHPLT dataset, we obtain encoder embeddings for 1000 randomly sampled occurrences per target word from HPLT v3.0 T5

monolingual models and the XLM-R model (Conneau et al., 2020); we additionally produce encoder embeddings from 100 randomly sampled occurrences per target word from HPLT v3.0 GPT-BERT (Charpentier and Samuel, 2024) monolingual models⁸ (Oepen et al., 2025).

4.2 Lexical substitutes

In addition to language model embeddings, we also consider lexical substitutes as a different kind of contextualized representations. Substitutes-based semantic change quantification methods were shown to do well at both LSCD benchmarks (Card, 2023; Periti et al., 2024) and downstream tasks like semantic change discovery (Umarova et al., 2025).

While it is possible to perform masked language modelling with T5 models, the results are not suitable for lexical substitutions generation, since these models were pre-trained with the span masking objective and tend to predict longer sequences rather than single lexemes. Examples can be found in Appendix C. For this reason, we use the HPLT v3.0 GPT-BERT models in a way similar to Card (2023) and Umarova et al. (2025) to represent 100 randomly sampled occurrences of each target word via top-15 substitutes. More details can be found in Appendix D. We also release XLM-R lexical substitutions. The number of target words for which we provide XLM-R embeddings and substitutions is limited by the size of the intersections between XLM-R and HPLT T5 tokenizer vocabularies. HPLT v3.0 GPT-BERT models use exactly the same tokenizers as the corresponding HPLT T5 models, so this issue is not relevant for them.

4.3 Word type embeddings

Although approaches based on *contextualized token embeddings* (as the ones described above) are the ‘daily drivers’ of modern LSCD researchers (Periti and Montanelli, 2024), we also publish *static type embedding* models trained on the DHPLT corpora. Static word embedding (SWE) models yield one vector representation per word type, as opposed to ‘a representation for each word occurrence’ from contextualized models. They were the LSCD mainstream until around 2021-2022 (Schlechtweg et al., 2020) and are still often used because of their simplicity and relatively modest compute requirements, both for training and for inference.

We train a SWE model for each language/time period combination using the SGNS architecture, also known as word2vec (Mikolov et al., 2013). For simplicity, we mostly use training hyperparameters from Aida and Bollegala (2025): window size 10, 5 epochs, 5 negative samples. Our embedding size is set to 300, and the model vocabularies are limited to 50000 most frequent words. Before training, the DHPLT documents are filtered to remove punctuation, as well as leading and trailing tabs and whitespaces.

Finally, for each language, the vector spaces of the models trained on time periods 1 (2011-2015) and 2 (2020-2021) are *aligned* to the vector space of the model trained on time period 3 (2024-), so as to make it possible to directly compute similarities between word embeddings in different models. We do this with the standard Procrustes alignment technique (Hamilton et al., 2016).

4.4 Frequency counts

Finally, we also publish frequency counts of each target word across the three DHPLT time periods. Changes in frequency of word usage coupled with lemma information are some of the very first indicators of changes in the word usage. These counts can also be used to control for frequency effects when quantifying semantic change (Card, 2023) and for planning compute usage (e.g. when generating lexical substitutions, half of the time is used for finding samples of the 100 least frequent target words, following the Zipf’s law (Powers, 1998)).

5 Sanity check

To demonstrate the utility of the DHPLT diachronic corpora, we look at the SWEs of the English word ‘AI’ (‘artificial intelligence’). Table 4 shows how the semantics of this term drifted. Back in the beginning of 2010s, it was associated almost exclusively with ‘AI characters’ in video games. A decade later, in 2020-2021, AI starts to be associated with ‘chatbots’ and machine learning, but it is still very much about robots, unmanned vehicles and Internet of Things. Only in the very last period 3 (2024-) we see the much too familiar landscape of LLMs, ChatGPT and ‘generative AI’. Interestingly, this trajectory is clear even despite the fact that (as noted above), our corpus for period 2 is bound to contain some documents created during the period 1, and the corpus for period 3 surely contains some documents from both 1 and 2.

⁸<https://hf.co/collections/HPLT/hplt-30-gpt-bert-models>

We observe a similar pattern (Table 5) when looking at the SWEs of the Spanish equivalent of the word: ‘IA’ (‘inteligencia artificial’). In 2010s, the word often appears in the context of gaming: ‘jugabilidad’, ‘PS’, ‘BETA’, etc. Then in 2020-2021, we start seeing words like ‘algoritmos’ and ‘tecnología(s)’ among closest semantic neighbours. These words follow more semantically-similar English-words such as ‘AI’, ‘artificial’, ‘learning’, etc, which are likely to still appear in the Spanish DHPLT corpora as part of names and titles. Finally, in 2024 ‘IA’ starts being associated with ‘generativa’ and ‘ChatGPT’. Very similar trends are found in our SWEs trained on Russian DHPLT documents (Table 6).

For T5 encoder embeddings, we calculate average pairwise distance between different time period representations (Kutuzov et al., 2022b) for the English lemmas ‘ai’, ‘remote’, ‘legislative’, and ‘jurisdiction’. The exact scores are to be found in the Appendix E. The change of ‘ai’ semantics is the largest and corresponds to the aforementioned SWE findings, while the changes of ‘legislative’, and ‘jurisdiction’, which are terms from the conservative legal domain, are the smallest. The degree of ‘remote’ change is somewhere in between and is the largest between 2011-2015 and 2020-2021 (‘Covid’) periods, when this word began to refer specifically to remote work rather than other contexts⁹. These observations hold for Spanish DHPLT corpora as well (Table 3).

6 Conclusion

We present DHPLT (‘Diachronic HPLT’): an open collection of large-scale diachronic corpora in 41 languages of 12 different language families. It is based on the web-crawled HPLT v3.0 datasets (Oepen et al., 2025), using web crawl timestamp as the temporal signal. The collection covers three time periods: 2011-2015, 2020-2021 and 2024. We augment DHPLT with pre-computed token-level semantic representations for language-specific sets of target words, to make it easier for practitioners to start experimenting with our corpora. Finally, we provide aligned static (type-based) word embedding models for each language and time period.

DHPLT (partially) addresses the lack of multilingual diachronic corpora in the LSCD field. We hope it will help making the landscape of historical

⁹<https://languages.oup.com/word-of-the-year/2020/>

language change modelling more rich and diverse. It should be especially relevant for studies in semantic change discovery.

Limitations

The main limitation of DHPLT is the source of temporal signal: that is, web crawl timestamps. As described above, timestamp X on a document does not guarantee that the text in this document was not created in the time periods $< X$ (earlier than X). It guarantees only that the text was not created in the time periods $> X$ (later than X). This difference compared to traditional diachronic corpora should be kept in mind when working with DHPLT.

Another limitation is that we provide only some of the possible types of semantic representations, and only for selected sets of target word, not for *all* the words in each of our languages. This is inevitable, given compute and storage space constraints. DHPLT allows practitioners to come up with their own sets of target words, or even conducts semantic change discovery experiments on the entire corpus. Our representations were obtained from models that do not incorporate any temporality, thus, e.g. masked language modeling predictions for 2011-2015 can include proper names that in fact first emerged in later periods. For future work, one might employ approach from (Fittschen et al., 2025) and pre-train models on texts from corresponding periods only.

Finally, in this paper we only introduce the DHPLT dataset; we leave conducting full-scale semantic change discovery on it for future work.

Acknowledgments

The computations were performed on resources provided by Sigma2 - the National Infrastructure for High-Performance Computing and Data Storage in Norway. This work was also in part supported by a gift from Google. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Google.

References

Taichi Aida and Danushka Bollegala. 2025. [SCD-Tour: Embedding axis ordering and merging for interpretable semantic change detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14775–14785, Suzhou, China. Association for Computational Linguistics.

- Pierpaolo Basile, Giovanni Semeraro, and Annalina Caputo. 2019. [Kronos-it: a dataset for the Italian semantic change detection task](#). In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 423–428, Bari, Italy. CEUR Workshop Proceedings.
- Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joonas Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. [An expanded massive multilingual dataset for high-performance language technologies \(HPLT\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.
- Dallas Card. 2023. [Substitution-based Semantic Change Detection using Contextual Embeddings](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association for Computational Linguistics.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024. [GPT or BERT: why not both?](#) In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.
- Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. [ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection](#). In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024a. [Definition generation for lexical semantic change detection](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.
- Mariia Fedorova, Timothee Mickus, Niko Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024b. [AXOLOTL’24 shared task on multilingual explainable semantic change modeling](#). In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 72–91, Bangkok, Thailand. Association for Computational Linguistics.
- Elisabeth Fittschen, Sabrina Li, Tom Lippincott, Leshem Choshen, and Craig Messner. 2025. [Pre-training language models for diachronic linguistic change discovery](#). *Preprint*, arXiv:2504.05523.
- Markus Freitag and Yaser Al-Onaizan. 2017. [Beam search strategies for neural machine translation](#). In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.
- Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarov. 2022. [Do not fire the linguist: Grammatical profiles help language models detect semantic change](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 54–67, Dublin, Ireland. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [Diachronic word embeddings reveal statistical laws of semantic change](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Sinan Kurtyigit, Maïke Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. [Lexical semantic change discovery](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998, Online. Association for Computational Linguistics.
- Andrey Kutuzov and Lidia Pivovarov. 2021. [Three-part diachronic semantic change dataset for Russian](#). In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.
- Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittmann. 2022a. [Nor-DiaChange: Diachronic semantic change dataset for Norwegian](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.
- Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022b. [Contextualized embeddings for semantic change detection: Lessons learned](#). In *Northern European Journal of Language Technology, Volume 8*.
- Zhidong Ling, Taichi Aida, Teruaki Oka, and Mamoru Komachi. 2023. [Construction of evaluation dataset for Japanese lexical semantic change detection](#). In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 125–136, Hong Kong, China. Association for Computational Linguistics.

- Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. [CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages](#). In *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia. Institute of Contemporary History.
- Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.
- Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitya Malik, Vladislav Mikhailov, and 13 others. 2025. [HPLT 3.0: Very large-scale multilingual resources for LLM and MT. Mono- and bilingual data, multilingual evaluation, and pre-trained models](#). *Preprint*, arXiv:2511.01066.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [FineWeb2: One pipeline to scale them all – adapting pre-training data processing to every language](#). *Preprint*, arXiv:2506.20920.
- Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. [Analyzing semantic change through lexical replacements](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand. Association for Computational Linguistics.
- Francesco Periti and Stefano Montanelli. 2024. [Lexical semantic change through large language models: a survey](#). *ACM Computing Surveys*, 56(11):1–38.
- Francesco Periti and Nina Tahmasebi. 2024. [A systematic comparison of contextualized word embeddings for lexical semantic change](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.
- Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. [GASC: Genre-aware semantic change for Ancient Greek](#). In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics.
- David M. W. Powers. 1998. [Applications and explanations of Zipf’s law](#). In *New Methods in Language Processing and Computational Natural Language Learning*.
- Marko Pranjic, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2024. Tracking semantic change in Slovene: A novel dataset and optimal transport-based distance. *arXiv preprint arXiv:2402.16596*.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [SemEval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.
- Khonzoda Umarova, Lillian Lee, and Laerdon Kim. 2025. [Current semantic-change quantification methods struggle with discovery in the wild](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35342–35355, Suzhou, China. Association for Computational Linguistics.
- Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. [LSCDiscovery: A shared task on semantic change discovery and detection in Spanish](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

A DHPLT datasets

Figure 1 shows the number of documents from different crawls for English and Georgian languages in HPLT v3.0.

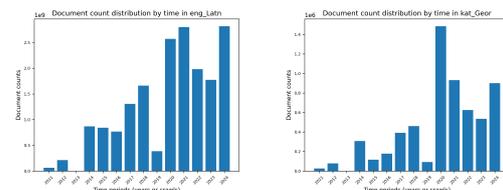


Figure 1: Number of documents per crawl year in the HPLT v3.0 datasets: English (left) and Georgian (right).

DHPLT files contain one document per line, with the following data fields:

- **id**: unique document identifier, can be used to link back to the original HPLT dataset,
- **ts**: timestamp, the exact date and time when the document was crawled from the Web,
- **text**: the actual document body, split into ‘segments’ (in most cases equal to paragraphs) with line break symbols,
- **doc_scores**: a list of integer ‘quality scores’ assigned to each segment of the document; to produce these scores, HPLT project employs heuristics from Web Docs Scorer (WDS).¹⁰

Table 1 lists all the DHPLT languages and their statistics.

B DHPLT target words

Below we provide technical details on selection of the target words.

For a given language L , we start from the vocabulary V_{T5_L} of the corresponding T5 model from the HPLT v3 T5 model collection. These models were pre-trained on documents in specific languages from the original HPLT v3.0 dataset. We assume that words which don’t appear as their own token in the corresponding T5 vocabulary are not frequent enough, so we omit them.

Further, we exclude tokens from V_{T5_L} that are word pieces or non-words. To do this, we first employ a heuristic for identifying full words. For most languages after removing punctuation we split the documents from diachronic corpora by whitespace, and count occurrences of such ‘full words’. For languages where splitting by whitespace doesn’t make sense, we use specific splitters. For Japanese, we employ `fugashi`¹¹ library that utilizes dictionaries for tokenization. For simplified Chinese, we segment words in text using `jieba`¹² library. Finally, for Thai, we go with word tokenization via `pythainlp`¹³ library.

Next, we count occurrences of such ‘full words’ across the diachronic corpora and filter out infrequent terms using a minimum frequency threshold. Thus, we only keep a token from V_{T5_L} if it appears

Language	ISO Code	Family	1	2	3
Albanian	als_Latn	Indo-European	0.5M	1M	1M
Arabic	arb_Arab	Afro-Asiatic	1M	1M	1M
Bosnian	bos_Latn	Indo-European	1M	1M	1M
Bulgarian	bul_Cyrl	Indo-European	1M	1M	1M
Catalan	cat_Latn	Indo-European	1M	1M	1M
Czech	ces_Latn	Indo-European	1M	1M	1M
Chinese	cmn_Hans	Sino-Tibetan	1M	1M	1M
Danish	dan_Latn	Indo-European	1M	1M	1M
German	deu_Latn	Indo-European	1M	1M	1M
Estonian	ekk_Latn	Uralic	0.5M	1M	1M
Greek	ell_Grek	Indo-European	1M	1M	1M
English	eng_Latn	Indo-European	1M	1M	1M
Finnish	fin_Latn	Uralic	1M	1M	1M
French	fra_Latn	Indo-European	1M	1M	1M
Hebrew	heb_Hebr	Afro-Asiatic	1M	1M	1M
Croatian	hrv_Latn	Indo-European	1M	1M	1M
Hungarian	hun_Latn	Uralic	1M	1M	1M
Armenian	hye_Armn	Indo-European	0.5M	1M	0.5M
Indonesian	ind_Latn	Austronesian	1M	1M	1M
Italian	ita_Latn	Indo-European	1M	1M	1M
Japanese	jpn_Jpan	Japanese	1M	1M	1M
Georgian	kat_Geor	Kartvelian	0.5M	1M	0.5M
Korean	kor_Hang	Korean	1M	1M	1M
Lithuanian	lit_Latn	Indo-European	1M	1M	1M
Latvian	lvs_Latn	Indo-European	0.5M	1M	1M
Macedonian	mkd_Cyrl	Indo-European	0.5M	1M	1M
Dutch	nld_Latn	Indo-European	1M	1M	1M
Norwegian	nob_Latn	Indo-European	1M	1M	1M
Polish	pol_Latn	Indo-European	1M	1M	1M
Portuguese	por_Latn	Indo-European	1M	1M	1M
Romanian	ron_Latn	Indo-European	1M	1M	1M
Russian	rus_Cyrl	Indo-European	1M	1M	1M
Slovak	slk_Latn	Indo-European	1M	1M	1M
Slovenian	slv_Latn	Indo-European	1M	1M	1M
Spanish	spa_Latn	Indo-European	1M	1M	1M
Swedish	swe_Latn	Indo-European	1M	1M	1M
Tamil	tam_Taml	Dravidian	0.5M	1M	1M
Thai	tha_Thai	Tai-Kadai	1M	1M	1M
Turkish	tur_Latn	Altaic	1M	1M	1M
Ukrainian	ukr_Cyrl	Indo-European	1M	1M	1M
Vietnamese	vie_Latn	Austro-Asiatic	1M	1M	1M

Table 1: DHPLT languages, writing systems, language families and historical period sizes (in millions of documents).

as a ‘full word’ at least 10 times in each of the three time periods (i.e., at least 30 times across all diachronic corpora in that language). Note that we ignore case when counting frequencies: e.g., if ‘operation’ and ‘OPERATION’ each occurs 5 times in the corpus, we keep both.

Further, following [Kurdyigit et al. \(2021\)](#), we also limit target words to only nouns, verbs, or adjectives. We use Stanza ([Qi et al., 2020](#)) part-of-speech taggers for all languages except Macedonian, for which we use `classla` ([Ljubešić et al., 2024](#)). In cases where the tagger identifies a token as a proper noun but its lower-cased version is tagged as a noun, the ‘NOUN’ tag takes precedence. Additionally, with the exception of Chinese, Japanese and Korean, we remove single character

¹⁰<https://github.com/pablo16n/web-docs-scorer>

¹¹<https://github.com/polm/fugashi>

¹²<https://github.com/fxsjy/jieba>

¹³<https://github.com/PyThaiNLP/pythainlp>

tokens in all languages.

Finally, we ensure that words in the target set are written in this language’s main script. Following the approach similar to `alphabet-detector`¹⁴, we use the Unicode names of characters in a target word candidate to verify its script. For instance, for English, we check that each character in each word from the target set is from Latin Unicode blocks, while for Japanese, we check that characters are either Hiragana, Katakana, or Kanji. We exclude words in which at least one character doesn’t belong to the expected script.

Figure 2 shows the distribution of target word counts across languages in DHPLT.

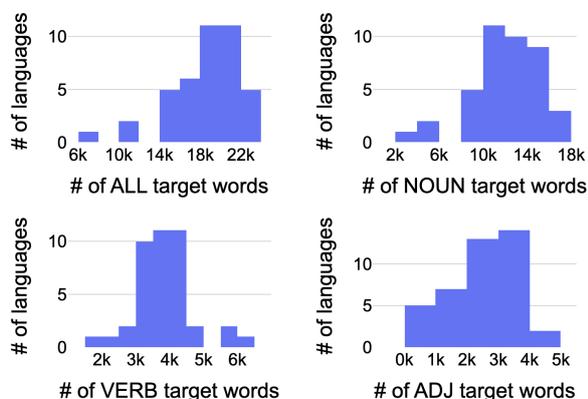


Figure 2: Number of target words across 41 languages for all target words (top left), target words that are nouns (top right), verbs (bottom left), and adjectives (bottom right).

C T5 substitutions

We use the same input format in our experiments, as shown in T5 model cards¹⁵. The reasons for not using T5 to generate lexical substitutions are :

- both encoder and decoder parts are required, duplicating the compute usage
- while it is technically possible to obtain not only the one most probable, but also top-k predictions with beam search (Freitag and Al-Onaizan, 2017), in practice the difference between predictions is too vague to make them useful. For example, the input *I remember we [MASK_1] it but still - do we use the decoder*

¹⁴<https://github.com/EliFinkelshteyn/alphabet-detector>

¹⁵<https://huggingface.co/collections/HPLT/hplt-30-t5-models>

part because the final hidden states of the encoder part are not mapped to vocabulary logits? (with the word ‘discussed’ behind the mask) yields the predictions ‘used the decoder part to do’, ‘used the decoder part for’, and ‘used the decoder part before’. Such small differences in generated representations are rather a problem for LSCD than an advantage, as shown in Fedorova et al. (2024a).

- predictions do not necessarily represent the semantics of the target word and tend to repeat other terms from the sentence, which is also observed on the aforementioned example
- longer inputs from real-world HPLT documents yield even longer predictions, for example, *How do you know the apples you are using for hard cider are ripe? Maybe, you would [MASK_1] me to define ripe. Is ripe defined by the ideal time to harvest an apple, to eat an apple, or to press an apple. We could even consider the question of ripeness for cooking apples. In my ... Continue reading When are apples ripe?* yields ‘say that the apples you are using for hard cider are ripe. But I don’t think that is the right way for’

D BERT representations

HPLT 3.0 T5s and HPLT 3.0 GPT-BERTs use the same tokenizer vocabulary for each language, so each target word has representations produced by both models.

E Examples

E.1 Sanity check for T5 embeddings

Tables 2 and 3 show change degrees of the English (‘ai’, ‘remote’, ‘legislative’, ‘jurisdiction’) and Spanish (‘ia’, ‘remoto’, ‘legislativo’, ‘jurisdicción’) words, according to the average pairwise distance, APD method (Fedorova et al., 2024a), on their respective T5 token embeddings.

Period pairs	‘ai’	‘remote’	‘legislative’	‘jurisdiction’
1 to 2	0.5533	0.4586	0.4117	0.4495
1 to 3	0.5646	0.4619	0.4141	0.4497
2 to 3	0.48	0.4548	0.4191	0.4351

Table 2: Average pairwise distances for several English target words calculated on T5 encoder embeddings.

Period pairs	‘ia’	‘remoto’	‘legislativo’	‘jurisdicción’
1 to 2	0.5733	0.5104	0.4031	0.4470
1 to 3	0.5763	0.4955	0.3925	0.4438
2 to 3	0.5810	0.4821	0.3979	0.4423

Table 3: Average pairwise distances for several Spanish target words calculated on T5 encoder embeddings.

E.2 Sanity check of HPLT 3.0 GPT-BERT substitutions

We perform manual analysis of the same 4 English words as in Section 5. The observations obtained with SWE models still hold: in 2011-2015, the words predicted as substitutions for ‘ai’ are either non-technical, or related to games or cars. In 2020-2021, a wider range of technologies is mentioned, including ‘IoT’, ‘NLP’, ‘robotics’, ‘animation’, etc. Also a lot of terms reflecting the social influence of AI emerge: ‘cybersecurity’, ‘humanity’, ‘innovation’, names of states and companies. Finally, in 2024, the trend of discussing social consequences continues: ‘elite’, ‘censorship’, ‘communism’, ‘scammers’, ‘capitalism’; much less technical terms and much more human-related ones are observed. There are also mentions of spheres which traditionally were human-dominated but has become automated recently: ‘art’, ‘healthcare’ etc. Surprisingly, we don’t observe many ‘LLM’-related terms among GPT-BERT’s predictions, but rather a shift from the optimistic perception of AI to the pessimistic one.

In 2011-2015, ‘remote’ is associated with networks and being spatially (geographically) distant. In 2020-2021, ‘virtual’ frequently occurs. In 2024, the associations show a techno-optimistic pattern similar to that of ‘AI’ in 2020-2021: positive job-related adjectives (‘skilled’, ‘flexible’, ‘professional’), wider range of technologies and spheres (‘satellite’ and ‘healthcare’ emerge). We also see terms related to society: state names, ‘climate’, ‘rural’.

Substitutions of ‘jurisdiction’ and ‘legislative’ bring no surprises, being related to law throughout all three time periods.

To conclude, representations obtained from contextualized models are sensitive to particular contexts at prediction, and thus capture more fine-grained semantic nuances than SWE models.

E.3 Sanity check for SWEs

Table 4 shows the semantic trajectory of the English word ‘AI’ in the DHPLT time periods, accord-

ing to our static word embedding models (SWEs). Similarly, Table 5 shows the trajectory for ‘IA’ that stands for ‘inteligencia artificial’ in Spanish across the three DHPLT time periods. Table 6 does the same for the Russian abbreviation ‘ИИ’ (‘AI’) (the model trained on the first time period does have this word in its vocabulary).

1: 2011-2015	2: 2020-2021	3: 2024-
multiplayer	chatbots	generative
NPCs	IoT	AI’s
RPG	robotics	GenAI
animations	RPA	ChatGPT
FPS	intelligence	LLMs

Table 4: Top 5 nearest neighbours (by cosine similarity) of the English term ‘AI’ in DHPLT static word embedding models by time periods. Case is ignored.

1: 2011-2015	2: 2020-2021	3: 2024-
BETA	AI	generativa
PS	artificial	artificial
AI	algoritmos	AI
jugabilidad	learning	inteligencia
artificial	inteligencia	ChatGPT

Table 5: Top 5 nearest neighbours (by cosine similarity) of the Spanish term ‘IA’ in DHPLT static word embedding models by time periods. Case is ignored.

2: 2020-2021	3: 2024-
интеллект (intellect)	интеллект (intellect)
AI	нейросети (neural networks)
роботов (robots)	ChatGPT
блокчейн (blockchain)	AI
алгоритмы (algorithms)	искусственный (artificial)

Table 6: Top 5 nearest neighbours (by cosine similarity) of the Russian term ‘ИИ’ (‘AI’) in DHPLT static word embedding models by time periods. Case is ignored. The 2011-2015 model does not have the word in its vocabulary (because of low frequency in this time period).

Transparent Semantic Change Detection with Dependency-Based Profiles

Bach Phan-Tat¹, Kris Heylen^{1,2}, Dirk Geeraerts¹, Stefano De Pascale^{1,3}, Dirk Spielman¹

¹Department of Linguistics, KU Leuven

²Instituut voor de Nederlandse Taal

³Vrije Universiteit Brussel

Correspondence: ttbach.phan@kuleuven.be

Abstract

Most modern computational approaches to lexical semantic change detection (LSC) rely on embedding-based distributional word representations with neural networks. Despite the strong performance on LSC benchmarks, they are often opaque. We investigate an alternative method which relies purely on dependency co-occurrence patterns of words. We demonstrate that it is effective for semantic change detection and even outperforms a number of distributional semantic models. We provide an in-depth quantitative and qualitative analysis of the predictions, showing that they are plausible and interpretable.

1 Introduction

Lexical semantic change detection is a well-established field in Natural Language Processing (NLP), with several shared tasks for different languages, amongst which the most popular one is SemEval 2020 Task 1 (Schlechtweg et al., 2020). Most of the models use either static word embeddings like word2vec (Mikolov et al., 2013) or ELMo (Peters et al., 2018) or contextualised ones such as BERT (Devlin et al., 2019) or S-BERT (Reimers and Gurevych, 2019) which is the current state-of-the-art (Cassotti et al., 2023). Despite the differences in their architectures, they all build upon the distributional semantics hypothesis (Firth, 1957) and all indirectly represent meanings as numerical vectors, abstracting away from the level of individual co-occurring items. Yet, linguists have long observed that we can study semantics and word-usage by directly looking at the untransformed co-occurrence contexts of the target (McEnery et al., 2019; Stubbs, 1995), which suggests that we could study semantic change by looking at shifts in the contexts of the target. However, most co-occurrence methods, including a recent one by McEnery et al. (2019), rely on surface co-occurrences (i.e., bag-of-words approaches to

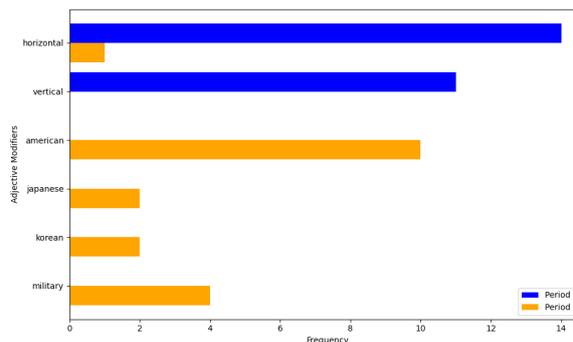


Figure 1: Changes in the frequencies of 6 adjectival modifiers of the English noun *plane* between the 2 sub-corpora for English of the SemEval 2020 shared task 1 (Schlechtweg et al., 2020). *Plane* is annotated as semantically changed in the dataset

co-occurrence). Surface co-occurrence often suffers from accidental and/or indirect co-occurrences and the arbitrary choice of the span size (Evert, 2008). Syntactic (dependency) co-occurrence (Evert, 2008; Garcia and García Salido, 2019; Seretan, 2011), with its different dependency relations, is more helpful to separate the signals. Consider the English noun *plane*: in the 20th century, the AIRCRAFT meaning emerged and became more dominant with the rise of such adjectival modifiers as *american*, *korean*, *japanese*, *military* whereas the DIMENSIONAL meaning became less dominant with the decrease of such adjectival modifiers as *horizontal*, *vertical*, as shown in Figure 1.

Using distributional shifts of the dependency co-occurrences to detect lexical semantic change is the focus of this paper. We investigate to what extent we could use explicit linguistic information to detect semantic change. Our main hypothesis is that the degree of semantic change of a word is reflected in changes in its dependency co-occurrences. Note that we are not trying to set a new state-of-the-art but to show the usefulness of explicit linguistic information and more simple, transparent techniques.

We investigate this question experimentally¹ using the dataset for English, German, Swedish, Latin from SemEval 2020 Task 1. Our main findings are the following:

1. Tracing the changes in the distribution of dependency co-occurrences outperforms a number of distributional models.
2. As our method is transparent by design, we can easily understand how a word has changed (as in the *plane* example above). In contrast, dense embedding models often require post-hoc interpretability methods or additional analyses (e.g., Integrated Gradients (Sundararajan et al., 2017), LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017), see surveys in Madsen et al., 2022; Arrieta et al., 2019) to obtain comparable, feature-level explanations. This makes our approach particularly convenient for linguistic and humanities studies where the explanation needs to be directly inspectable.
3. Our approach is compatible with linguistic theories of lexical semantics and semantic change (e.g., Geeraerts et al., 2023; Geeraerts, 1997; Hanks, 2013; Dalpanagioti, 2019), allowing researchers to bridge the gap between NLP and diachronic linguistics.

2 Related work

A long tradition in distributional semantic models treat lexical meaning as a function of usage: words are characterised by the contexts they appear in, and semantic change is operationalised as a shift in these contextual distributions over time. Within this tradition, work roughly splits into two dominant methodological styles.

The first tracks change directly in interpretable co-occurrence evidence (e.g., shifts in salient collocates or in the distribution of context features). This line has strong roots in mid-20th-century structuralism, combining Harris’s distributional analysis (Harris, 1954) with Firth’s contextual theory of meaning (Firth, 1957). With the rise of corpus linguistics and concordancing, this idea became a practical methodology. Corpus linguists compared a target word’s prominent collocates and recurring patterns across corpora and later across time

slices, often supported by association measures for identifying characteristic collocations (Church and Hanks, 1990; Sinclair, 1991). This collocational tradition also developed notions such as semantic preference/prosody, which are systematic attitudinal or functional tendencies inferred from a word’s typical collocational environments (Partington, 1998; Stubbs, 1995). This naturally lends itself to diachronic analysis when those environments shift. This is done by building time-sliced collocation profiles and examining how association-ranked collocates emerge, fade, or are replaced across periods, keeping the evidence directly inspectable. Case studies have used such collocation turnover to document long-term changes in conventionalised combinations (e.g., Alba-Salas, 2007; Pettersson-Traba, 2021). Some of the more recent semi-automatic methods were developed by Garcia and García Salido (2019) and McEnery et al. (2019). Different tools have also made this style scalable, enabling extraction and comparison of diachronic collocation phenomena in large corpora such as AntConc (Anthony, 2014), #Lancsbox (Brezina et al., 2021), SketchEngine (Kilgarriff et al., 2014), DiaCollo (Jurish, 2015; Jurish and Nieländer, 2020).

The second maps co-occurrence evidence into vector-space representations (count-based or neural embeddings). The first models used an untransformed co-occurrence matrix (later changed to other measures such as Mutual Information (MI) or Log Likelihood Ratio (LLR)), with each row corresponding to a word and the columns words in the vocabulary (Bullinaria and Levy, 2007; Gulordava and Baroni, 2011; Schütze, 1998). The result is that each word will be represented by a sparse count vector. Predictive neural embedding architectures like word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017) could be said to capture similar information as the count-based approach but are more powerful at abstracting information from the corpus with denser vectors (Baroni et al., 2014). Contextualised models such as BERT (Devlin et al., 2019) are even more powerful in the sense that they can distinguish between individual instances of word usages instead of representing one word as one vector (Tahmasebi and Dubossarsky, 2023). Several attempts have been made to integrate explicit linguistic information into neural network models, such as Dependency Skip-gram (Levy and Gold-

¹All the codes used in this paper are available at <https://github.com/phnatatbach/LChange26-Dep>

berg, 2014) or SynGCN and SemGCN (Vashishth et al., 2019). Semantic similarities between words or tokens (in the case of contextualised models) are often calculated using cosine distance between their vectors. For semantic change detection tasks, the standard workflow is as followed (Tahmasebi and Dubossarsky, 2023; Tahmasebi et al., 2021):

1. Data are split into different periods;
2. Semantic vectors of the target words are trained for each time period,;
3. Different embedding spaces are aligned;
4. Semantic changes are quantified by comparing aligned representations, often via cosine distance between word vectors, or via clustering comparison of representations across periods.

While collocation-profile approaches keep the evidence directly interpretable, they often yield descriptive outputs that are harder to compare systematically across datasets and languages. As lexical semantic change (LSC) research moved toward standardised evaluation with different standard shared tasks and dataset, amongst which the most popular one is SemEval 2020 Task 1 (Schlechtweg et al., 2020), methods that produce compact numeric representations and a single comparable distance score became increasingly attractive. As a result, the majority of works and even the current state-of-the-art are all applications of this line of method (e.g., Cassotti et al., 2023; Geeraerts et al., 2023; Giulianelli et al., 2020; Kim et al., 2014; Kulkarni et al., 2015; Kutuzov et al., 2018; Montariol et al., 2021; Tahmasebi and Dubossarsky, 2023; Tahmasebi et al., 2021).

There are a few exceptions to this trend, who did not rely on dense numerical vectors, such as Ryzhova et al. (2021) and Kutuzov et al. (2021) who used the distributions of grammatical features to detect semantic changes and achieved great results, even surpassing many embedding methods in some tasks and datasets; or Tang et al. (2023) who compared sense distributions to detect semantic change. Inspired by their work, we want to investigate the ability of using only explicit linguistic information (in this case, syntactic co-occurrences) in capturing word meaning change.

Most of the work in collocation/co-occurrence analysis uses bag-of-words window co-occurrence

and suffers from arbitrariness in their choice of context windows and from accidental co-occurrences (Evert, 2008). We instead rely on co-occurrence analysis with dependency defined contexts. Rather than using surface window contexts, it treats dependency relations (slots) as interpretable dimensions, so predicted change could be attributed to concrete shifts in those dependency environments. This is conceptually similar to the DepDM model of the Distributional Memory Framework (Baroni and Lenci, 2010) and to the Sketch Engine’s Word Sketch Difference (Kilgarriff et al., 2014) but in our case, we keep the raw dependency relations instead of merging them, as will be discussed below. Also, most of them only focus on the salient co-occurrences with such measures as logDice (Rychlý, 2008), MI (Church and Hanks, 1990) or Local MI (Baroni and Lenci, 2010), and LLR (Dunning, 1993), thus do not measure the dynamic of the shift of the co-occurrence profile. We quantify semantic change using Jensen-Shannon Divergence (Menéndez et al., 1997). It is calculated between the distribution of the lexical fillers (slot fillers) of each slot across periods. We evaluate our method in the standard shared tasks, SemEval 2020 Task 1 (Schlechtweg et al., 2020) with 4 languages. Because we use all dependency slots, our method is unsupervised, making it directly comparable to other systems. Finally, we also conduct in-depth qualitative analyses to highlight both the usefulness and limitations of our approach.

3 Data and tasks

Following the SemEval-2020 task on Unsupervised Lexical Semantic Change Detection, we formulate the task as either binary classification or graded ranking, corresponding to Subtask 1 and Subtask 2 in SemEval-2020 (Schlechtweg et al., 2020). In Subtask 1, given a set of target words, a system has to predict whether each word gained or lost at least one sense between two time periods. In Subtask 2, a system has to rank target words by their degrees of semantic change. Annotating data for semantic change is non-trivial since it requires judging meaning differences across many usage instances from each period. A widely adopted annotation protocol is DUREl (Schlechtweg et al., 2018), where annotators are asked to compare pairs of sentences containing the target word and judge whether the word is used in the same or a different sense, then these pairwise judgments are aggregated into a final

change score, either binary or continuous. This was the procedure used in the creation of the SemEval 2020 dataset. It covers four languages, namely English (37 targets), German (48), Latin (40), and Swedish (32), with manual annotations for both subtasks. Each language is accompanied by a diachronic corpus split into two time periods. Using that dataset allows us to directly compare our approach against state-of-the-art distributional systems under a shared evaluation setup.

4 Method

4.1 Data preprocessing

Because our method requires dependency relations, which are not available in the original dataset, we first parsed the data using Stanza (Qi et al., 2020) given its high performance on different languages. Upon inspection, we noticed some discrepancies between the Stanza outputs and the lemmatised SemEval data in **part-of-speech (POS) tags** and **lemmas**.

POS mismatches The original dataset did not take into account the POS tags of the target lemmas except for English where targets are provided as lemma_POS combinations (e.g., `attack_nn` but not `attack_vb`) because annotation was performed for specific lemma-POS pairs (Schlechtweg et al., 2020). However, because our package requires the POS tags as part of the input, ensuring POS consistency is crucial. We found cases where tokens corresponding to nominal targets were tagged as PROPEN by Stanza (causing missed target occurrences if we filter by NOUN), and cases where adjectival targets were tagged as VERB when used as participles.

Lemma mismatch We also found cases where Stanza’s lemmatization diverges from the official target lemma due to orthographic variation and OCR errors. One notable example is the German noun *Lyzeum*, which is consistently lemmatised as *Lyceum* by Stanza in the earlier period (old standard spelling), leading to systematic mismatches with the SemEval target form.

To quantify the impact of these discrepancies, we prepared two parallel versions of the parsed data, Stanza-faithful (SZ) and SemEval-faithful (SE). In the SZ version for all languages, we only do minimal normalisation to the target lemmas, namely changing PROPEN to NOUN and spelling variations. In the SE version of German, Latin, Swedish, we

ensure that the Stanza results match those of the original lemmatised dataset by replacing all POS tags of the target lemmas with TAR and normalising the spelling variations of the target lemmas. For the English SE version, we split the dataset into multiple sub-datasets based on the official lemmatised files where the targets are marked. Each sub-dataset corresponds to one target lemma.

4.2 Basic procedure

For the main experiment, we used SynFlow (Phan-Tat, 2025), a custom package developed specifically for this method.

For each target lemma (e.g., the English noun *plane*), we extract all of its dependency slots (e.g., adjective modifiers or `amod`) and the slot-fillers and their POS tags (e.g., `new/A`, `old/A`, `vertical/A`, `horizontal/A`). Note that because the CoNLL-U file does not encode the directionality in the `deprel` column but only in the HEAD indices, we have to use the prefix `chi_` (children) if the slot-filler depends on the target and the prefix `pa_` (parent) if the target depends on the slot-filler. We preserve distinct dependency labels instead of aggregating semantically similar roles (e.g., combining `nsubj:pass` and `obj`). Doing so has two advantages. First, this semantic granularity allows us to detect specific structural shifts, such as an increase in passive voice usage, which would otherwise be obscured by a coarse-grained object category. Second, it is generalisable across languages while merging has to be performed on a case-by-case basis and requires language specific knowledge.

For each target word and for each of its slots, we now have a different distribution of their slot-fillers for each period. To quantify how this distribution changes across periods, we use Jensen-Shannon Divergence (JSD). JSD is a symmetric, smoothed variant of Kullback–Leibler divergence (Kullback and Leibler, 1951) and when computed with log base 2, it is bounded between 0 and 1, making the scores easy to interpret and comparable across slots. Moreover, JSD can be decomposed into per-filler contributions, enabling fine-grained attribution of which fillers drive a slot’s change. By contrast, prior work on Grammatical Profiles (Kutuzov et al., 2021; Ryzhova et al., 2021) measured change with cosine distance between time-specific frequency vectors, which compares vector geometry (angles) rather than directly comparing probability distribu-

tions.

After computing slot-level JSD scores for each target lemma, we aggregate them into a single lemma-level change score by taking the mean JSD over slots whose JSD exceeds 0.5. We use a mean rather than a sum because lemmas differ in how many slots they have and summing would systematically inflate scores for lemmas with more slots. Using the maximum is also undesirable because 1) very rare slots with small sample size can yield extreme JSD values (often 1), which would make many lemmas appear equally ‘maximally changed’ and 2) of our assumption that semantic change is manifested in multiple slots. There is little established guidance on choosing an appropriate JSD threshold, and sparsity becomes more severe when distributions are split by slot, which can push JSD values upward compared to larger distributions. The 0.5 value was empirically selected (based on our previous experience applying JSD to sparse linguistic data) to balance sensitivity to genuine linguistic change against robustness to noise. It serves as a high-pass filter for semantic salience, which isolates substantial structural shifts while suppressing minor fluctuations inherent to sparse distributions. Note that we intentionally avoid any supervised finetuning (e.g., for selecting the optimal JSD cut-off) to preserve the fully unsupervised protocol and comparability with other unsupervised approaches.

To sum up, the inputs to our method are dependency-parsed corpora and a list of target lemmas. SynFlow then

1. Collects the slot fillers for each dependency slot;
2. Computes the JSD scores for every slot;
3. Aggregates the slot-level JSDs greater than 0.5 into a lemma-level JSD score.

The output is a dictionary of the form {target lemma: aggregate JSD score}. In the next sections, we describe 2 improvements to this basic procedure.

4.3 Frequency filtering

To reduce noise that could be introduced by rare items, we exclude slot-fillers that appear only once in all periods before calculating the JSD.

4.4 POS removal

In the basic procedure above, we initially retain POS tags for slot fillers. However, because our analysis is driven primarily by dependency relations, and because POS tags can introduce additional noise (e.g., adjective vs. participial verb ambiguities as discussed above), we remove the POS information from slot-fillers before computing JSD. This reduces the risk that parser-induced POS inconsistencies create spurious differences between periods and ensures that divergence reflects changes in dependency defined co-occurrence rather than tagging artefacts.

5 Results

We evaluate our method on both subtasks of the SemEval 2020 Unsupervised Lexical Semantic Change Detection shared task. As described in Section 3, Subtask 1 is a binary classification task and is evaluated based on accuracy. Subtask 2 is a ranking task and is evaluated with Spearman’s rank correlation. The main focus of our method would be on **subtask 2**. We then use our subtask 2 scores for subtask 1 classification, following the strategy of [Kutuzov et al. \(2021\)](#).

5.1 Subtask 2

The results for subtask 2 are reported in table 1. For a full comparison with SemEval-2020 Task 1 participating systems, see Table 5 in [Schlechtweg et al. \(2020\)](#).

Basic procedure The slot-level JSD aggregation described above yields an average Spearman correlation of 0.096 on the SE dataset and 0.108 on the SZ dataset. This is below the shared-task baseline, but still higher than some systems based on type and token embeddings (e.g., UoB, TUE).

Frequency filtering Removing rare slot-fillers (Section 4.3) improves the performance substantially, especially on the SE dataset: the average correlation across languages nearly doubles, and Latin improves by an order of magnitude (0.026 → 0.215). On the SZ dataset, the effect is mixed: English drops to −0.001, while the other three languages improve.

POS Removal Removing POS tags from slot-fillers yields only a small additional gain: +0.021 on SE and +0.005 on SZ on average.

System	Average	English	German	Latin	Swedish
Basic Procedure					
SE	0.096	0.108	0.135	0.026	0.115
SZ	0.108	0.139	0.120	0.017	0.155
Frequency Filtering					
SE 2	0.229	0.266	0.253	0.215	0.181
SZ 2	0.129	-0.010	0.191	0.142	0.192
POS Removal					
SE No POS	0.117	0.115	0.175	0.038	0.141
SZ No POS	0.113	0.139	0.145	0.001	0.166
Frequency Filtering + POS Removal					
SE 2 No POS	0.239	0.277	0.258	0.258	0.162
SZ 2 No POS	0.139	-0.008	0.193	0.197	0.173
Prior SemEval results					
Count baseline	0.144	0.022	0.216	0.359	-0.022
Best shared task system (Cassotti et al., 2023)	0.583	0.757	0.877	-0.056	0.754

Table 1: Performance in graded change detection (SemEval 2020 Subtask 2), Spearman rank correlation coefficients. SE = SemEval, SZ = Stanza.

System	Average	English	German	Latin	Swedish
Basic Procedure					
SE	0.580	0.595	0.604	0.475	0.645
SZ	0.583	0.541	0.604	0.475	0.71
Frequency Filtering					
SE 2	0.593	0.649	0.604	0.475	0.645
SZ 2	0.569	0.486	0.604	0.475	0.71
POS Removal					
SE No POS	0.59	0.595	0.646	0.475	0.645
SZ No POS	0.583	0.541	0.604	0.475	0.71
Frequency Filtering + POS Removal					
SE 2 No POS	0.606	0.649	0.604	0.525	0.645
SE 2 No POS DP	0.585	0.621	0.604	0.5	0.613
SZ 2 No POS	0.565	0.486	0.604	0.525	0.645
Prior SemEval results					
Count baseline	0.613	0.595	0.688	0.525	0.645
Best shared task system	0.687	0.622	0.75	0.7	0.677

Table 2: Performance in binary change detection (SemEval 2020 Subtask 1), accuracy. Note that in this paper we mostly focus on ranking (Subtask 2). All the binary change detection methods here are entirely based on the scores produced by the ranking methods. SE = SemEval, SZ = Stanza.

Frequency Filtering + POS Removal Combining both strategies produces our strongest overall results. Performance improves for all languages except Swedish (where the combined setting reaches 0.162, slightly below our Swedish best of 0.192). The combined setting on the SE dataset achieves an average Spearman correlation of 0.239, which is nearly double the Count baseline and higher than

a number of embedding-based systems. This is impressive given the simplicity of the approach.

5.2 Subtask 1

The results for subtask 1 are presented in table 2. We follow the thresholding strategy used by Kutuzov et al. (2021) by assigning a classification score of 1 to the top 43% of the target words for

each language. The ranking is obtained from their JSD scores as discussed above. The different average accuracy scores from the different systems are tightly clustered (between 0.58 and 0.606) and are very close to the accuracy of the baseline, which is already surprisingly hard to beat.

Our best system is still SE 2 No POS (i.e., SemEval + frequency filtering + POS removal), which achieves an average accuracy of 0.606, showing that simple denoising techniques (i.e., frequency filtering and POS removal) work. In the case of Swedish, 3 of our system variants even outperform the state-of-the-art, achieving an accuracy of 0.71. We also tested our best system with dynamic programming² (Truong et al., 2020) and observed that it does not cause any significant accuracy decrease when compared to the hard-coded 43% cut-off point. This indicates that our method does not require knowledge of the test data distribution

6 Qualitative analysis

The strong point of our system is its interpretability, in the sense that it allows the users to see which kinds of shift drive the change through JSD decomposition, and check the concordances for the locus of semantic change. In this section, we analyse the prediction of the system qualitatively to understand what is being captured.

Based on the rankings of the Subtask 2 Gold Scores and our system predictions, we divided the target lemmas into five groups: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) and MID. A lemma is TP if it falls in the top 33% of both the gold and predicted rankings, and TN if it falls in the bottom 33% of both. A lemma is FP if it is in the bottom 33% of the gold ranking but the top 33% of our method’s ranking, and FN if it is in the top 33% of the gold ranking but the bottom 33% of our method’s ranking. All remaining lemmas are assigned to the MID group. We focus our qualitative analysis on TP, TN, FP, and FN, as these categories most clearly reveal our method’s strengths and limitations. Due to space constraints, in the main text we present one illustrative example per group and language (the most extreme cases). The lemmas of each group are listed in Table 3 in the Appendix.

²We applied the offline change point detection algorithm on the sorted JSD scores. The method identifies a split point k that minimises the total within-segment squared error: $SSE(y_{1:k}) + SSE(y_{k+1:n})$, where $SSE = \sum y^2 - \frac{(\sum y)^2}{n}$.

Also, in the analysis, we will only look at slots with JSD of at least 0.5 as that is the cut-off point used in the previous procedure. A detailed description of the dependency slots presented in this section is provided in Table 4 in the Appendix.

True positives Upon qualitative analysis, our method does indeed capture relevant co-occurrence information related to semantic change. The English noun *graft* develops a new MEDICAL use (transplanting tissue) and a new POLITICAL use (political corruption) in the later period, whereas in the earlier period it is used primarily in agricultural contexts. This is reflected in 3 out of 5 dependency slots, such as governing verbs (pa_obj, e.g., *cut the graft*) and compound modifiers (chi_compound, e.g., *skin graft*) as shown in Figure 2. In the first period, fillers and contexts predominantly relate to cultivation and horticulture (e.g., *adapting the graft to the soil*, *flowering graft*), while the second period is characterised by an increase of medical and political contexts (e.g., *help the skin graft*, *government graft*). The remaining two slots show weaker or less interpretable shifts. The German adjective *abgebriiht* has developed a new FIGURATIVE meaning (e.g., *abgebriihter Manager* ‘a hardened manager’) from the LITERAL one (e.g., *abgebriihter Reis* ‘blanched rice’), as evidenced in the modified nouns (pa_amod) and modifying adverbs (chi_advmod). The Latin noun *pontifex* shifts from ‘state-religious officer’ toward ‘the Pope’. This is reflected in 13/15 slots. In the adjectival modifier (chi_amod) and oblique nominals (pa_obl) slots, for example, *pontifex* co-occurs less with administrative vocabulary (e.g., *gero* ‘manage’) and

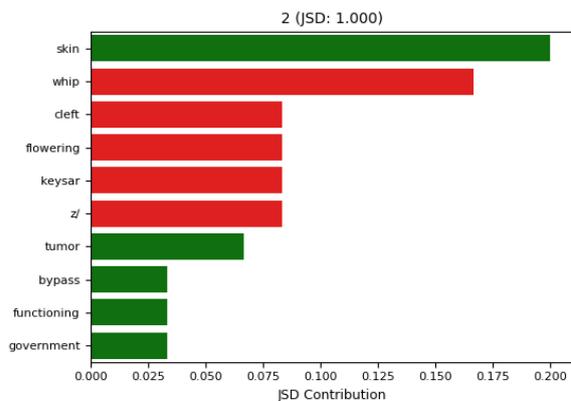


Figure 2: The JSD contributions of different slot-fillers in the slot *chi_compound* of the English noun *graft*. Green bars indicate an increase in relative frequency; red bars indicate a decrease.

more with Christian contexts (e.g., *summus pontifex* ‘Supreme Pontiff’, *oro* ‘pray’). Finally, the Swedish noun *uppläggnings* has developed a new abstract STRUCTURING meaning from the physical STOCKPILING meaning, with consistent evidence in all of its slots. For example, in the nominal modifier slot (pa_obl), there is an increase in *räkenskap* ‘accounting’ while in the earlier period it is dominated by types of goods such as *jernplätars* ‘of iron plates’, *spannmåls* ‘of cereals’.

True negatives Regarding TNs, we observe that few dependency slots exceed the JSD threshold and even when they do, the margin is often negligible. This is the case for the English noun *tree*. Our method detects a moderate shift in nominal modifiers (chi_nmod, JSD = 0.517) and compounds (JSD = 0.533), but these shifts are driven by a transition from biological attributes (e.g., *trees of this size*) to locative contexts (e.g., *trees of this yard*). Crucially, however, both filler sets remain semantically consistent with the core PLANT sense, indicating mere topical variation rather than genuine semantic evolution. Even if there are slots with high divergence (e.g., ≈ 0.8), the aggregate lemma-level score remains low because other stable slots dilute the overall mean. This was the case for the German verb *vergönnen*, the Latin adjective *necessarius* and the Swedish verb *uträtta*.

False positives FPs are target words for which our method detects large shifts in one or more slots, yet the lemma is ranked low in the gold Subtask 2 ranking. A common pattern is slot-level turnover driven by a small number of highly characteristic contexts, which inflates divergence without corresponding to a clear sense innovation under the annotation scheme. For the English noun *face*, the high change score is driven primarily by the two compound slots (pa_compound, chi_compound). The divergence is dominated by a few fixed expressions, most notably *face value* and product-related compounds such as *face cream* and *face lift*. This illustrates the limitation where the substantial frequency change of a small set of expressions can cause the method to overestimate the semantic change. The Swedish noun *kokärt* is also stable in meaning, yet all of its slots exhibit high JSD due to contextual turnover. The German noun *Seminar* presents an interesting case. We observe a metonymic shift: INSTITUTIONAL DEPARTMENT \rightarrow COURSES. This is strongly evidenced by a high divergence (0.926) in

the nominal heads (pa_nmod), transitioning from institution-related fillers (e.g., *Lehrer*, *Director*, *Übungsschule*) to participant-oriented ones (e.g., *Teilnehmer*). However, its official gold label for Subtask 1 is 0 (i.e., no sense gain or loss) and its official Subtask 2 rank is only 39/48 (i.e., low degree of semantic change). This suggests our system successfully detected an overlooked semantic shift in the official annotations. The Latin noun *dolus* is another interesting case. It is officially classified as 1 in Subtask 1 and our method successfully identified the word’s narrowing and pejoration from TRICK (neutral) to FRAUD (negative) in the governing verb (pa_obj) slot (JSD = 0.933). However, its official ranking in Subtask 2 is low (29/40). This strengthens the organisers’ finding: models that excel at one subtask are not necessarily good at the other (Schlechtweg et al., 2020).

False negatives FNs are target words whose aggregate JSD scores underestimate the degree of change with respect to the gold ranking. We attribute a substantial portion of these errors to our slot-selection threshold. Although the English noun *record* develops a salient BEST PERFORMANCE sense (e.g., *world record*), reflected by a sharp rise of *set* and *break* in governing verbs (pa_obj), the divergence for this slot is 0.48, just below our 0.5 cutoff. This signal is thus excluded from aggregation, causing the lemma-level score to miss the underlying shift. A similar threshold effect appears for the Swedish noun *motiv*, which shifts from a predominantly REASON meaning (typically framed in moral evaluation) to a predominantly ARTISTIC SUBJECT meaning (often framed in aesthetic evaluation) between the two periods. While detectable in adjectival modifiers (chi_amod), the slot’s low JSD (0.436) leads to its exclusion, again resulting in an underestimated aggregate score. In other cases, the change signal is present but remains only moderately strong at the slot level, so it does not lift the aggregate score enough to match the gold ranking. The German verb *ausspannen* shows a clear metaphorisation: PHYSICAL (unharnessing/stretching out) \rightarrow RELAX (taking time off). This is visible in its direct objects (chi_obj) with an increase in temporal fillers (*Tag*, *Stunde*) and a decrease in physical objects (*Draht*, *Faden*), or in the passive subject (chi_nsubj:pass). However, their JSDs are only 0.543 and 0.584, respectively. This illustrates that genuine semantic innovations can yield relatively modest divergences in individ-

ual slots, resulting in underestimated lemma-level change. A related pattern occurs for the Latin noun *potestas*, where broadening (LEGAL AUTHORITY → GENERAL CAPABILITY) is reflected in two low JSD slots, adjectival modifiers (*chi_amod*, JSD = 0.587) and nominal modifiers (*chi_nmod*, JSD = 0.643).

7 Conclusion

This paper demonstrates that explicit linguistic co-occurrence evidence, when structured through syntax, is sufficient to build an effective lexical semantic change detection system. Our method achieves competitive performance across languages and can even outperform several distributional models in some tasks and languages. Although our method does not match state-of-the-art systems based on large pretrained language models, which can encode a wider range of linguistic and contextual information, maximising performance metrics was not our primary objective. Instead, we want to bring more attention to theory-driven and interpretable approaches.

To further refine this approach, future work should:

1. Replace hard slot thresholds with more adaptive, evidence-weighted aggregation to increase reliability;
2. Further develop theory-driven ways of selecting and merging slots to reduce non-essential variations (recall the *nsubj:pass* and *obj* example in 4.2), but note the cost of generalisability;
3. Explore hybrid methods that preserve slot-level attribution while using embeddings to stabilise sparse evidence (e.g., clustering slot-fillers using general type embeddings before computing JSD);
4. Take into account covariations between slots and looking at multiple slots simultaneously (e.g., to detect changes in valency/subcategorisation patterns)

More broadly, we view our method as an instance of a general framework for semantic change analysis: identify interpretable dimensions of a target’s distribution (here, dependency relations) and quantify change within each dimension before aggregating into a global measure. This framework

could (and should) be used as a complementary tool to interpret and validate the changes detected by other approaches.

8 Limitations

We highlight three primary limitations and practical considerations of our proposed method.

First, dependency availability is a hard constraint, making this method inapplicable to low-resource languages lacking robust dependency parsers. Even for high-resource languages, pre-processing choices matter: POS and lemma mismatches between gold targets and parsers (e.g., Stanza) output requires alignments.

Second, sensitivity to noise (e.g., rare slots, rare slot-fillers) remains a challenge. Denoising (e.g., filtering singletons and removing POS tags from fillers) proved central to our performance.

Third, our qualitative analysis, while supporting the interpretability claim, simultaneously reveals inherent constraints in our scoring mechanism. The method can *overestimate* change when a small number of fixed expressions or highly productive co-occurrences dominate slot turnover (as seen with idioms). Conversely, it can *underestimate* change when genuine semantic innovations fall just below the heuristic slot-selection cutoff.

Acknowledgments

This project has received funding from the European Union’s Horizon Europe programme for research and innovation under MSCA Doctoral Networks 2022, Grant Agreement No. 101120349 and Grant Agreement No. 101119511.

We would like to thank the anonymous reviewers for their detailed and informative comments and suggestions.

References

- Josep Alba-Salas. 2007. [On the life and death of a collocation: A corpus-based diachronic study of *dar miedo/hacer miedo*-type structures in Spanish](#). *Diachronica*, 24(2):207–252. Publisher: John Benjamins Type: Journal Article.
- Laurence Anthony. 2014. AntConc (Version 3.4.3)[Computer Software]. Tokyo, Japan: Waseda University.
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénézet, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and

- Francisco Herrera. 2019. [Explainable artificial intelligence \(xai\): Concepts, taxonomies, opportunities and challenges toward responsible ai](#).
- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. [Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.
- Marco Baroni and Alessandro Lenci. 2010. [Distributional memory: A general framework for corpus-based semantics](#). *Computational Linguistics*, 36(4):673–721.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching Word Vectors with Subword Information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146. Place: Cambridge, MA Publisher: MIT Press.
- Vaclav Brezina, Pauline Weill-Tessier, and Tony McEnery. 2021. [#lancsbox v. 6.x](#). Lancaster University. Software version 6.x.
- John A. Bullinaria and Joseph P. Levy. 2007. [Extracting semantic representations from word co-occurrence statistics: A computational study](#). *Behavior Research Methods*, 39(3):510–526.
- Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. [XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.
- Kenneth Ward Church and Patrick Hanks. 1990. [Word Association Norms, Mutual Information, and Lexicography](#). *Computational Linguistics*, 16(1):22–29.
- Thomai Dalpanagioti. 2019. [From corpus usages to cognitively informed dictionary senses](#). *Lexicography*, 6(2):75–104.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ted Dunning. 1993. [Accurate Methods for the Statistics of Surprise and Coincidence](#). *Computational Linguistics*, 19(1):61–74. Place: Cambridge, MA Publisher: MIT Press.
- Stephanie Evert. 2008. [Corpora and collocations](#). In Kytö M Lüdeling A, editor, *Corpus Linguistics. An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin, New York.
- John Rupert Firth. 1957. [A synopsis of linguistic theory 1930–1955](#). In *Studies in Linguistic Analysis*, pages 1–32. Basil Blackwell, Oxford.
- Marcos Garcia and Marcos García Salido. 2019. [A Method to Automatically Identify Diachronic Variation in Collocations](#). In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 71–80, Florence, Italy. Association for Computational Linguistics.
- Dirk Geeraerts. 1997. [Diachronic Prototype Semantics: A Contribution Historical Lexicology](#). Oxford University Press.
- Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang. 2023. [Lexical Variation and Change: A Distributional Semantic Approach](#), 1 edition. Oxford University Press.
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. [Analysing Lexical Semantic Change with Contextualised Word Representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
- Kristina Gulordava and Marco Baroni. 2011. [A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus](#). In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.
- Patrick Hanks. 2013. [Lexical Analysis: Norms and Exploitations](#). The MIT Press.
- Zellig S. Harris. 1954. [Distributional Structure](#). *WORD*, 10(2-3):146–162. Publisher: Routledge.
- Bryan Jurish. 2015. [DiaCollo: On the trail of diachronic collocations](#). In *CLARIN Annual Conference 2015 (Wrocław, Poland, October 15–17 2015)*, pages 28–31.
- Bryan Jurish and Maret Nieländer. 2020. [Using DiaCollo for Historical Research](#). In *Selected Papers from the CLARIN Annual Conference 2019*, pages 33–40.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. [The Sketch Engine: ten years on](#). *Lexicography*, 1:7–36. Publisher: Springer.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. [Temporal Analysis of Language through Neural Language Models](#). In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages

- 61–65, Baltimore, MD, USA. Association for Computational Linguistics.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. [Statistically Significant Detection of Linguistic Change](#). In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, pages 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. Event-place: Florence, Italy.
- S. Kullback and R. A. Leibler. 1951. [On Information and Sufficiency](#). *The Annals of Mathematical Statistics*, 22(1):79–86.
- Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. 2021. [Grammatical Profiling for Semantic Change Detection](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 423–434, Online. Association for Computational Linguistics.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. [Diachronic word embeddings and semantic shifts: a survey](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Omer Levy and Yoav Goldberg. 2014. [Dependency-based word embeddings](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Neural Information Processing Systems*.
- Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. [Post-hoc interpretability for neural nlp: A survey](#). *ACM Computing Surveys*, 55(8):1–42.
- Tony McEnery, Vaclav Brezina, and Helen Baker. 2019. [Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse](#). *International Journal of Corpus Linguistics*, pages 413–444.
- M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. 1997. [The Jensen-Shannon divergence](#). *Journal of the Franklin Institute*, 334(2):307–318.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient Estimation of Word Representations in Vector Space](#). ArXiv:1301.3781 [cs].
- Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. [Scalable and Interpretable Semantic Change Detection](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.
- Alan Partington. 1998. *Patterns and Meanings*. John Benjamins.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global Vectors for Word Representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep Contextualized Word Representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Daniela Pettersson-Traba. 2021. [Measuring semantic distance across time](#). *Journal of Research Design and Statistics in Linguistics and Communication Science*, 6(2):138–165.
- Bach Phan-Tat. 2025. [Synflow](#).
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python Natural Language Processing Toolkit for Many Human Languages](#). ArXiv:2003.07082 [cs].
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. [“why should I trust you?”: Explaining the predictions of any classifier](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.
- Pavel Rychlý. 2008. A Lexicographer-Friendly Association Score. *Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN*, 2:6–9. Publisher: Masaryk University.
- Anastasiia Ryzhova, Daria Ryzhova, and Ilya Sochenkov. 2021. Detection of semantic changes in russian nouns with distributional models and grammatical features. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021”*.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

- Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. [Diachronic Usage Relatedness \(DURel\): A Framework for the Annotation of Lexical Semantic Change](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.
- Hinrich Schütze. 1998. [Automatic Word Sense Discrimination](#). *Computational Linguistics*, 24(1):97–123. Place: Cambridge, MA Publisher: MIT Press.
- V. Seretan. 2011. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer Netherlands.
- John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.
- Michael Stubbs. 1995. [Collocations and semantic profiles: On the cause of the trouble with quantitative studies](#). *Functions of Language*, 2(1):23–55. Publisher: John Benjamins Type: Journal Article.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. [Axiomatic attribution for deep networks](#). In *International Conference on Machine Learning*.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. [Survey of computational approaches to lexical semantic change detection](#). In *Computational approaches to semantic change*. Zenodo.
- Nina Tahmasebi and Haim Dubossarsky. 2023. [Computational modeling of semantic change](#).
- Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. [Can word sense distribution detect semantic changes of words?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.
- Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. [Selective review of offline change point detection methods](#). *Signal Processing*, 167:107299.
- Shikhar Vashishth, Manik Bhandari, Prateek Yadav, Piyush Rai, Chiranjib Bhattacharyya, and Partha Talukdar. 2019. [Incorporating syntactic and semantic information in word embeddings using graph convolutional networks](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3308–3318, Florence, Italy. Association for Computational Linguistics.

Appendix

A Predictions of our method

	English	German	Latin	Swedish
True Positive	graft (noun) plane (noun) head (noun) bit (noun) player (noun)	abgebrüht Engpaß Ohrwurm Abgesang Eintagsfliege Dynamik Armenhaus	pontifex sacramentum scriptura titulus imperator sanctus	uppläggning krita medium konduktör uppfattning beredning notis
True Negative	tree (noun) risk (noun) multitude (noun) fiction (noun) contemplation (noun) bag (noun)	vergönnen Tier Ackgerät Frechheit aufrechterhalten	necessarius consilium hostis simplex voluntas poena	uträtta annandag studie
False Positive	face (noun) quilt (noun) lane (noun) savage (noun)	Seminar Truppenteil Mulatte Pachtzins Naturschönheit	dolus fidelis consul ancilla acerbus	kokärt vegetation
False Negative	record (noun) ounce (noun) rag (noun)	ausspannen artikulieren verbauen	potestas licet salus virtus	motiv ledning bearbeta

Table 3: True Positive, True Negative, False Positive, False Negative results of SE 2 NOPOS.

B Descriptions of the slots

Slot Name	Explanation
chi_advmod	The target is modified by the adverbial slot filler.
chi_amod	The target is modified by the adjectival slot-filler.
chi_case	The target is governing the prepositional slot-filler.
chi_compound	The target is being modified by the slot-filler in the compound.
chi_nmod	The target is modified by the nominal slot-filler.
chi_nsubj:pass	The target is the verb governing the passive subject.
chi_obj	The target is the verb governing the object slot-filler.
pa_amod	The target is the adjective modifying the slot-filler.
pa_compound	The target modifies the slot-filler in the compound.
pa_nmod	The target is the nominal modifier of the slot-filler.
pa_obj	The target is the object of the verbal slot-filler.
paobl	The target is the oblique of the slot-filler.

Table 4: A detailed description of the dependency slots presented in section 6.

Semantic Change Characterization with LLMs using Rhetorics

Jader Martins Camboim de Sá^{1,2}, Jooyoung Lee^{1,3}, Cédric Pruski², Marcos Da Silveira²

¹FSTM - University of Luxembourg

2 place de l'Université, L-4365, Esch-sur-Alzette, Luxembourg,

²Luxembourg Institute of Science and Technology

5 avenue des Hauts-Fourneaux, L-4362, Esch-sur-Alzette, Luxembourg,

³Brown University

Providence, RI 02912, United States,

Correspondence: first.second@list.lu

Abstract

Languages continually evolve in response to societal events, resulting in new terms and semantic shifts. These changes have significant implications for computer applications, including automatic translation and chatbots, making it essential to characterize them accurately. The recent development of Large Language Models (LLMs) has notably advanced natural language understanding, particularly in sense inference and reasoning. In this paper, we investigate the potential of LLMs in characterizing three aspects of polysemy and semantic shift: dimension, relation, and orientation. We achieve this by combining the reasoning capabilities of LLMs with rhetorical devices and conducting an experimental assessment of our approach using newly created datasets. Our results highlight the effectiveness of LLMs in capturing and analyzing semantic shifts, providing valuable insights to improve computational linguistic applications.

1 Introduction

Language evolution remains a subject of fascination and inquiry across diverse disciplines, including neuroscience, psychology, philosophy, and computational linguistics. Despite this interdisciplinary interest, our understanding of language is often superficial, with much to uncover regarding its dynamics (Allan, 2013; Pinker, 2003). Among the many elements that shape language, a central aspect in understanding its dynamics is the nature of polysemy and how word meanings shift (Campbell, 2013). This evolution is particularly intriguing in computational linguistics, as it impacts applications such as automatic translation and chatbots (Camboim de Sá et al., 2024).

While humans can rapidly adapt to these shifts using contextual information and cognitive processes to grasp the senses of a sentence or a word, it is complex to provide comparable cultural knowl-

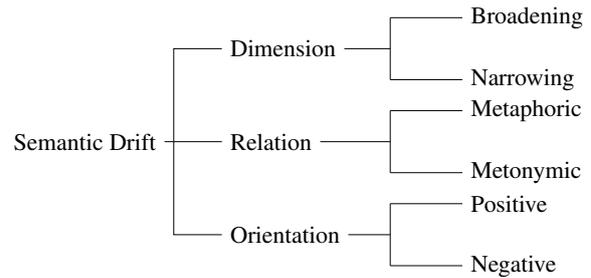


Figure 1: Taxonomy for the poles of Semantic Drift.

edge and nuance to machines. Consequently, machines often lack the tools to adapt to these variations and perform effective communication (Cambria et al., 2023). Therefore, in many modern Natural Language Processing (NLP) systems, we observe the impacts of prototypicality, polysemy, and semantic shift on end users, especially when a task requires deep contextual dependency (Camboim de Sá et al., 2024).

In contexts where historical or domain-specific knowledge of meaning is crucial, fields like Lexical Semantic Change Detection (LSCD) or Word Sense Disambiguation (WSD) have developed methods to detect these semantic shifts (Tahmasebi et al., 2018; Navigli, 2009). While a significant body of work explores which words have changed in different moments or domains, there is still a need for further comprehension regarding how these changes are associated and the implications of these shifts in these systems (Hengchen et al., 2021). For instance, in sentiment analysis, being aware that the term ‘sick’ has acquired a positive connotation could significantly alter the interpretation of a sentence.

Theories to comprehend semantic shift and polysemy exist in the domain of linguistics. One predominant typology (Traugott, 2017), categorizes shifts into broadening/narrowing (a word gains or loses senses), amelioration/pejoration (a

word is perceived more positively/negatively), and metaphorization/metonymization (the word is used as a metaphor or metonymy, respectively). We illustrate these types of shift in Table 1.

A child in dirty overalls.	
He used a dirty trick to win the competition.	pejoration
No other style of hat was acceptable with an evening dress.	
He took off his politician’s hat and talked frankly.	metaphorization
The diamond is currently set in the crown of the Queen.	
The colonies revolted against the crown .	metonymization

Table 1: Examples illustrating the characterization of types of shift.

In this paper, we address the problem of characterizing and analyzing how meaning drifts across contexts. Our methodology aims to instruct LLMs to utilize rhetorical devices and guided reasoning and characterize change within a comparative framework. Therefore, our contributions are:

- While zeugma was theoretically accepted as a method for sense differentiation, this study is the first quantitative inquiry of zeugma and other rhetorical devices as a diagnosis for semantics.
- A new approach to semantic change characterization exploring the “reasoning” and rhetorical capabilities of LLMs, not only improving model prediction but also rendering an interpretable analysis.
- We propose three new publicly available datasets designed to evaluate semantic change characterization across three aspects: dimension, orientation, and relation.

The paper is structured as follows: Section 2 presents related work in the field of semantic change characterization using LLMs. Section 3 presents the task solved in this paper. Section 4 details our methodology for prompting models for semantic change characterization. Section 5 introduces the experimental settings and results. Section 6 discusses insights from the method, while in Section 7 we discuss its limitations. Finally, Section 8 contains concluding remarks and outlines future work.

2 Related Work

Many papers on semantic change address the problem of identification, i.e., detecting if the meaning of a word has changed without inferring what type

of change occurred. State-of-the-art methods consist of encoding words as contextualized vectors and subsequently comparing them (Pivovarova and Kutuzov, 2021; Zamora-Reina et al., 2022).

In the context of Language Models (LMs), some authors explore semantic change identification as a sequence-to-sequence problem (Lyu et al., 2022; Giulianelli et al., 2023) by first prompting the model to disambiguate the word in context and then generating a contextualized word representation. Later, LLMs were employed for this task (Wang and Choi, 2023; Periti et al., 2024b), but using only a few-shot prompt or fine-tuning for the task.

For semantic change characterization, works apply restricted taxonomies (Camboim de Sá et al., 2024). In the context of broadening/narrowing, Bochkarev et al. (2022) utilize a neural network to determine if a word is employed as a named entity. This approach creates a temporal perspective of a word’s usage and allows them to compare occurrences of a word to see whether it has gained new usage in the corpus.

For metaphorization, Maudslay and Teufel (2022) fine-tuned a BERT model with supervision to classify contextualized words into metaphors and then analyzed different corpora. Finally, in the amelioration/pejoration context, Fonteyn and Manjavacas (2021) measure polarity in the term ‘to death’ by calculating the distance between the word vectors for ‘good’ and ‘bad.’

In the work of Cassotti et al. (2024), the authors explore a subset of Blank’s taxonomy (Blank et al., 2003) for synchronic change. They concatenate pairs of word definitions obtained from WordNet to train a classifier for predicting the type of semantic change (hyperonymy, hyponymy, co-hyponymy, antonymy, and homonymy) in a supervised manner. Similarly, in (Periti et al., 2024a), the authors also explore WordNet for types of semantic change (synonym, antonym, hypernym, and random), using distance embeddings for their analysis.

Compared to previous works on semantic change, this is the first study to use Chain-of-Thought (CoT) for this task, with our approach being deeply motivated by linguistic literature. In terms of semantic change characterization, this is the first work that generalizes across all types of change (Camboim de Sá et al., 2024), doesn’t require task-specific fine-tuning, and can be used for every type of relation, e.g., metaphor and metonymy.

3 Task Definition

In this paper, we propose a task for characterizing semantic change across different corpora. To this end, we rely on the following set of predominant types defined in the literature (Traugott, 2017; Campbell, 2013):

- **Broadening:** gaining a new meaning related or not to the previous meaning, such that a word represents more concepts, e.g., ‘cloud,’ a computing infrastructure.
- **Narrowing:** restriction of meaning occurs when a symbol represents fewer concepts than previously, e.g., ‘gay’ which historically meant festive or happy, is now predominantly used to refer to homosexuality only.
- **Amelioration:** a word gains a more positive sense compared to the previous sense, *nice*, ‘foolish, innocent’ changed to ‘pleasant.’
- **Pejoration:** the word is used with a worse connotation compared to the previous usage, *stincan*, ‘smell (sweet or bad)’ changed to *stink*.
- **Metonymization:** association between terms, e.g., *board* ‘table’, changed to “people sitting around a table, governing body.”
- **Metaphorization:** conceptualizing one thing in terms of another, e.g., ‘head of the company’ the word ‘head’ conceptualizes “command or control.”

While this typology was initially proposed for historical language change, this typology aligns with current interest and methodologies in computational corpus-based studies (Camboim de Sá et al., 2024) and synchronic typologies (Blank et al., 2003). Inspired by Camboim de Sá et al. (2024), we consider three poles, namely Dimension, Relation, and Orientation, given the addressed computational challenge (see Figure 1).

In the **dimension** pole, we compute the “number of senses” a word can have. This pole is self-complementary as increasing represents broadening, and decreasing represents a narrowing of senses. After identifying the number of senses, we can compare the differences between corpora.

Metaphorical and metonymical changes are classified under the **relation** category, as these changes enhance the connection between one sense of a

word and its other senses. In this framework, a word’s meaning relies on the link established through conceptual (abstract relation) or material (physical association) similarity between concepts. We identify which senses are used figuratively for other senses of the same word.

The **orientation** pole regroups the amelioration or pejoration of a meaning. In this pole, words are analyzed according to the contextual sentiment captured from each corpus, and then we analyze how the sentiment changes over corpora. In this study, we explore only positive, negative, and neutral sentiment values for words.

3.1 Dataset for Lexical Semantic Change Characterization

From a semasiological perspective, words’ meanings could be inferred from the context, for example, “He *targeted* me, after I didn’t agree with his proposal” or “The *mustache* guy, is coming today?”. This occurs because we use previous knowledge of the meaning and contextual information to infer the new meaning. Over time, if this usage becomes widespread enough, it transitions from creative expression to a conventional sense.

The optimal approach to LSCD should replicate the natural ability of humans to perceive and understand semantic change. This means employing unsupervised learning techniques, where the system infers the evolving sense of a word solely based on its prior exposure to the language and the contextual information within the data. In essence, the system learns to identify semantic shifts without needing pre-labeled data (Schlechtweg et al., 2020).

To evaluate our framework, we created, in collaboration with a native English-speaking linguist, an LSCD Characterization dataset following the Cambridge setting (Tang et al., 2013) as further detailed in the appendix E. The dataset is composed of pairs of sentences sharing the same word (see the example with the word ‘lost’ from equations (1) and (2)). The first sentence expresses one possible usage (e.g., original usage), while the second sentence expresses a different usage. The characterization task will detect the word’s meaning in each context and compare them to infer how it changed. To support this task, we created three new datasets, one for each pole of change, where the example pairs present the type of change the LLM should be able to characterize.

In the **dimension** dataset, we curated the WiC

data (Pilehvar and Camacho-Collados, 2019) to get a fraction of reliable and high-quality examples. The original dataset only classified the word’s meaning as related or unrelated. We adapted it according to the DUREL format (Schlechtweg et al., 2020). We define that the word’s meanings are identical if the same meaning is observed when we merge the sentences using zeugma without producing a bad pun (see the ‘lost’ example, equation (3)) and different otherwise.

We define words as related if they have a direct relation (metaphor or metonymy) for their usage. Otherwise, we define them as unrelated. For example, the word ‘head,’ with the sense of ‘control,’ can only be marked as related if it directly represents figurative usage from a literal sense, i.e., ‘head’ as ‘upper part of the body.’

For the **relation** data, we collected examples from the metaphor detection dataset (Choi et al., 2021) to get literal and metaphorical usages and examples from the literature to increase the evaluation dataset size. Sentences were manually collected using online dictionaries like Linguee¹ and Merriam-Webster² and verified by three experts. We performed similarly to collect sentences containing metonymies from both the literature (Lakoff and Johnson, 2008) and reliable Web sources.

Last, we created the **orientation** data by acquiring sense pairs for the same word where we had the most significant difference between negative and positive connotations from these pairs expressed in SentiWordNet (Baccianella et al., 2010). The sentence pairs were obtained from SemCor (Raganato et al., 2017) and WordNet (Miller, 1995) depending on how much context they have to infer the sense given the sentence. Additionally, we transform the sentences so that the sentiment of a word cannot be trivially detected from the whole sentence. Hence, the detector needs to comprehend the word-level sentiment. Sometimes, we modify the sentence to be negative while the word meaning is positive, e.g., ‘my life is a hell, **darling**.’

Table 2 contains the number of samples for each dataset and the accuracy of a random classifier.

4 Methodology

LLMs have exhibited significant progress in natural language comprehension. This includes reasoning by analogy (Webb et al., 2022), understand-

¹www.linguee.com

²www.merriam-webster.com

Task	Labels	Random	Total
Dimension	Identical , Different	0.53	260
Relation	Metaphor, Metonymy, Unrelated	0.35	331
Orientation	Positive, Negative , Neutral	0.29	262

Table 2: General view of the three datasets created for Lexical Semantic Change Characterization. ‘Random’ represents the accuracy of a random classifier for each dataset.

ing metaphors (Liu et al., 2022), argumentation (Chen et al., 2023), and acquiring cultural knowledge (Petroni et al., 2019). Additionally, instructing an LLM to generate a rationale, which is a natural language explanation for its reasoning process, before providing an answer has been shown to improve performance on many NLP tasks that require logical reasoning (Wei et al., 2022; Kavumba et al., 2023). This rationale generation step is believed to inject more information retrieved from the LLM’s internal knowledge store into the prompt. This enriched prompt allows the LLM to consider a broader range of knowledge during the final decision-making process (Dasgupta et al., 2022).

Building on these advances, we leverage the cultural knowledge embedded in LLMs to enhance contextual understanding for the task, positioning the LLM as an annotator within the framework outlined in the previous subsection. The problem of characterizing LSCD, identifying different senses, figurative usage, and feelings, relies on building cognitive relations between other senses that depend on human perception and culture (Lakoff and Johnson, 2008). Since language is the most effective way to explain itself (Pinker, 2003), our approach aims to replicate the cognitive process of Huang and Chang (2023) by using the human knowledge embedded in LLMs and rationales to generate the same associations that humans naturally make (Dasgupta et al., 2022; Strachan et al., 2024).

Tracking and comparing word senses is a complex task (Kilgarriff, 1997). To address this problem, we approach it as a “comparative semantics” problem, i.e., instead of extracting the meaning as a final objective, we rely on the idea of relatedness and likeness of meaning to compute LSCD. Similar to Schlechtweg et al. (2024), we compare word occurrences for the characterization problem. However, in DUREL, the annotator is prompted with two sentences that share a particular word, and the annotator must classify the similarity level between contextual senses. In our approach, we

use an LLM instead of humans to annotate the relatedness between words, and we reduce it to just ‘identical’ and ‘different’ classes.

We perform the characterization starting with a cognitive semasiological comparative analysis of the word meaning (Kilgarriff, 1997), following the Cambridge setting (Tang et al., 2013). We first provide a context where the word sense could be inferred (the Gracian approach (Agirre and Edmonds, 2007)), to then decide, via reasoning, if the deduced senses are different according to a class of change (Blank et al., 2003).

This step is done by an LLM that acts like a judge/annotator using a special type of CoT (Wei et al., 2022) with detailed step-by-step reasoning (Mitra et al., 2023) to elicit the model’s ability of word sense induction and comparative semantics. Our approach exploits rhetorical techniques to produce ‘cognitive-appealing’ arguments on how the senses differ.

For the **dimension** pole, we created a prompt requesting a word sense differentiation where the prompt asks if a word is used in an identical or different sense. Knowing how many unique senses exist, the task of measuring word dimension reduces to the problem of sense counting. To perform the sense differentiation, we instruct the model to use **zeugma** (Kearns, 2006) as a cognitive approach to identify identical senses. If it can produce a consistent zeugma, the senses are identical. Otherwise, the words are different.

Zeugma is a rhetorical device where a single word, typically a verb or an adjective, governs or modifies two or more words in a sentence. This device creates a clever or unexpected relationship between different sentence parts. Zeugma often results in a play on words contributing to the overall impact of the expression, and it adds a layer of complexity or humor to the language used in a sentence and allows us to explore the sense usage difference (Kearns, 2006). For example:

- (1) "He lost his keys."
- (2) "He lost his temper."
- (3) "He lost his keys and his temper." (?)

In sentence (3), the word “lost” is used to combine both sentences in a related sense to describe both (1) losing physical objects (keys) and (2) losing emotional control (temper). This zeugma creates a figurative and compact expression that links two different related meanings of the word ‘lost’

in a single sentence, creating a bad pun (Kearns, 2006). This bad pun arises because the second usage of ‘lost’ does not preserve the same sense as the first, indicating a difference in the meaning. In Figure 2, we present part of the prompt employing zeugma for the dimension dataset.

Sense Differentiation

[...]. Follow these steps to complete the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Write a sentence that joins both using zeugma and the same shared word while preserving the same sense. If the construction makes a bad pun, the words have a different sense.
- Step 4. Based on the previous reasoning, give your final answer: ‘identical’ or ‘different.’

[Few-shot examples.]

Figure 2: Prompt for sense differentiation in the dimension dataset. Complete prompts can be verified in Appendix D.

For the **relation** pole, we experiment with a prompt instruction for figurative language analysis. The instruction is based on a simplified version of the Metaphor Identification Procedure (MIP) (Steen et al., 2007). The model is designed to differentiate between metaphor and metonymy by establishing relationships between the underlying concepts. This relation can be either abstract, suggesting a metaphorical mapping between domains (evoking tropes), or material, if an association exists between the concepts.

- (1) "The main objective of this forthcoming decision will be preparation for the winter."
- (2) "Winter can cause many disruptions for public transport."
- (3) "The word ‘winter’ in the second sentence is associated with its problems, such as snow, making it a metonymy."

The provided examples showcase how **simile** acts as a parsing mechanism for the AI model. By leveraging similes, the model can reframe figurative language based on the similarity or association between concepts. In essence, a simile acts

as a rhetorical device that explicitly compares two entities to enhance the explanation and detail the nature of that comparison. By deciphering the figurative meaning within context, we aim to guide the model towards extracting more information about the underlying semantic relationship. This, in turn, allows the model to make a more accurate prediction regarding the type of relation – whether it’s a metaphor or a metonymy. In Figure 3, we illustrate the prompt for obtaining the figurative association between word usages.

Sense Figurativeness

[...]. Follow these steps to complete the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Compare the usage, determining if the second is related as a metaphor (where the word is used in a similar but non-literal sense), as a metonymy (where the word represents something closely related to or associated with it), or unrelated, used with a different sense.
- Step 4. Based on the third step, write the final answer, 'metaphor', 'metonymy', or 'unrelated.'

[Few-shot examples.]

Figure 3: Prompt for figurative sense in the relation dataset.

For the **orientation** pole, the current state-of-the-art sense-level sentiment analysis requires first a WSD step, then a sentiment analysis step (Zhang et al., 2023). Similar to the previous prompts and following the best practice, we instruct the model to perform a textual sense disambiguation and then differentiate the orientations between these senses (Wiebe and Mihalcea, 2006). To differentiate orientation, we use **antagonage** to compare senses’ positiveness (or negativeness) and to enrich contextual information on how these senses can be perceived in the training data.

Antanagoge is a rhetorical device that involves responding to an accusation or negative point with a counter-argument or positive point. It is used to mitigate the impact of something negative by placing it alongside something positive. We use the common sentence “I’d rather X than Y” as a few-shot demonstration to instruct the LLM to get the

most probable contextual ordering. The example below exemplifies the usage of antanagoge.

- (1) "A terrific presentation."
- (2) "A terrific storm."
- (3) "I'd rather have a terrific presentation than a terrific storm."

Sense Orientation

[...]. Follow this instructions to execute the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Leverage the rhetorical strategy of antanagoge, contrasting a negative with a positive, to weigh why one meaning might be more favorable than the other, or if they stand neutral.
- Step 4. Based on the third step, write the final answer 'negative', 'positive', or 'neutral.'

[Few-shot examples.]

Figure 4: Prompt for sense orientation in the orientation dataset.

5 Experiments

In this section, we present prompting details and measure the quality of our annotations for LSCD.

5.1 Experimental results

We compare rhetorics with two baselines to evaluate how good LLMs and rhetorical devices are for characterizing semantic change. The first baseline is based on few-shot prompting only. The second additionally asks to perform a CoT. All evaluated prompting methods have 3-shot examples with the correct label.

In Appendix D, we share the complete prompt for the rhetorical reasoning for all the three poles of change. We also provide the code for reproduction in Git³. We selected LLaMA-3 and Phi-3 as the current state-of-the-art LLMs for the instruction prompt. We sampled the models 5 times for each method with temperature $\tau = 0.7$, using the guidance⁴ library to control the generated layout.

³https://git.list.lu/d4h/change_characterization

⁴<https://github.com/guidance-ai/guidance/>

We report the mean and standard deviation of the accuracy.

In Table 3, we present the results for the dimension dataset. We can observe that the rhetorical method meaningfully improves the accuracy of Phi-3 and LLaMA-3-70b over the baselines, while for LLaMA-3-8b, the best method is the few-shot prompt. While the data used for instruction tuning LLaMA-3 is not publicly released, we believe it was fine-tuned on WiC data (Pilehvar and Camacho-Collados, 2019), which could explain the improved accuracy.

Method	LLaMA-3-8b	LLaMA-3-70b	Phi-3
3-Shot	.75±.00	.76±.00	.60±.00
CoT	.59±.02	.75±.02	.62±.01
Rhetorical (ours)	.68±.03	.78±.01	.71±.02

Table 3: Accuracy metrics for the Dimension dataset.

Table 4 shows our results for the relation dataset. For both models, the rhetorical method improved significantly over the baselines, with LLaMA-3-70b getting the overall best capabilities for recognizing figurative usage.

Method	LLaMA-3-8b	LLaMA-3-70b	Phi-3
3-Shot	.38±.00	.52±.00	.49±.00
CoT	.45±.01	.53±.01	.51±.03
Rhetorical (ours)	.52±.02	.59±.02	.56±.01

Table 4: Accuracy metrics for the Relation dataset.

In Table 5, the accuracy is tied between few-shot and rhetorical for LLaMA-3-8b, while for Phi-3 it is 10 points higher. This task is notably more complex than the others for LLMs. It’s expected as for humans it’s also a challenging task, we observed a Fleiss’ Kappa of 0.3104 and Krippendorff of 0.3115.

Method	LLaMA-3-8b	LLaMA-3-70b	Phi-3
3-Shot	.37±.00	.36±.00	.30±.00
CoT	.34±.03	.35±.00	.30±.02
Rhetorical (ours)	.37±.03	.39±.01	.40±.02

Table 5: Accuracy metrics for the Orientation dataset.

In Appendix C, we provide a correlation plot across models and prompting strategies; we can observe that the rhetoric method generates the lowest correlation between strategies. This evidence points out that the initial bias of the model for inferring change in meaning is very different from the current established methods in linguistic literature.

6 Discussion

In the boxes below, we present some cherry-picked cases of reasoning with rhetorical devices. We first present the compared pairs in bullet points, followed by the model’s answer.

Example #9 battle

- During the course of this **battle**, four Canadian soldiers were wounded.
- I call on him not to let the **battle** of hierarchies within the directorates-general prevent this from being implemented.

In the first sentence, ‘battle’ is used literally, referring to a fight or conflict. — In the second sentence, ‘battle’ is used metaphorically to represent a struggle or conflict between different groups or factions. — The word ‘battle’ is related to the first sentence as a metaphor.

A: metaphor

For the figurative prompt, the Phi-3 model enriched the prompt with a reasonable inference for the second sentence; with this information, the model could correctly relate the senses as a metaphorical variation.

Example #167 salt

- She was **salty** because of what happened yesterday.
- She **salts** her lectures with jokes.

In sentence 1, ‘salty’ is used to describe someone who is angry or upset. — In sentence 2, ‘salts’ is used to describe the act of adding flavor to something, in this case, a lecture. — I’d rather have a lecture with jokes than be angry or upset. The second sense is more positive.

A: positive

The orientation prompt forces an order between the senses, making comparing non-trivial perceptions of good and bad feasible. In the example above, the model is instructed to use antanagoge and then impersonate a preference over both scenarios; this allows it to classify the second sense as more positive correctly.

The illustrative examples demonstrate that

LLMs can retrieve stored cultural knowledge and enrich prompts, allowing in-context learning to utilize more information for decision-making. However, in some cases, they fail to manipulate senses to produce zeugma correctly, and sometimes, the conclusions are inconsistent with the reasoning. We have detailed these failure cases in Appendix A.

This observation aligns with well-documented phenomena: LLMs can hallucinate and generate incorrect reasoning even if they reach the correct answer (Ye and Durrett, 2022). Other failure cases may involve the leakage of evaluation data and the confusion of generalization with memorization, as it is difficult to verify whether an LLM-generated figurative usage explanation is novel or directly derived from training data.

Meaning is a fundamental open question in NLP. While LLMs can often replicate human-like behavior by relying heavily on form, they struggle with simpler tasks that require a basic understanding of meaning (Berglund et al., 2023). Understanding how models deal with meaning in controlled settings (such as comparing a word with itself) is crucial for enabling models to generalize beyond mere form.

While rhetorical devices are standard tools in linguistics, as they are interpretable and simple for analysis of meaning, our understanding of their effectiveness is still evolving. Using LLMs, we can better test theories and explore how humans relate meanings through language, further advancing our understanding.

7 Limitations

Our sentences were carefully selected to enable a concrete understanding of the word’s meaning in the context; if the method is applied in under-contextualized sentences, it can result in wrong sense disambiguation, leading to bad overall performance. We evaluated our approach on LLMs optimized for English; while the methodology applies to other languages, the quality of characterization may differ. LLMs can hallucinate when generating explanations, so the generated rationale may not reflect the correct cognitive likeness between senses even when it correctly characterizes the change.

8 Conclusion

In this paper, we leveraged cultural knowledge in LLMs to relate and characterize the word’s meaning change for different contexts. We experi-

mentally observed that rhetorical devices can help LLMs to mimic human meaning associations interpretably, enabling semantic change characterization.

LLMs have many rhetorical examples in their training data, which are not trivially detected. In future works, we expect to explore in depth whether LLMs can produce new rhetorical associations or if they are just reproducing their training data; this task can be investigated through non-conventionalized metaphor or circumstantial metonymy.

Another essential aspect to investigate is how to better prompt language models for figurative reasoning if they can perform better by providing all forms of metaphors (instrumentality, container, etc.) (Lakoff and Johnson, 2008) and metonymies (part-whole, cause-effect, etc.) (Lakoff and Johnson, 2008) as examples. We also expect to extend the DUREL approach to automatize the characterization of semantic change.

Acknowledgements

The research reported in this publication was supported by the Luxembourg National Research Fund (FNR), project D4H grant number PRIDE21/16758026

References

- Eneko Agirre and Philip Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.
- Keith Allan. 2013. *The Oxford handbook of the history of linguistics*. OUP Oxford.
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. [SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. [The reversal curse: Lms trained on "a is b" fail to learn "b is a"](#). *ArXiv*, abs/2309.12288.
- Andreas Blank et al. 2003. Polysemy in the lexicon and in discourse. *Trends in linguistics studies and monographs*, 142:267–296.
- Vladimir V. Bochkarev, Stanislav V. Khristoforov, Anna V. Shevlyakova, and Valery D. Solovyev. 2022.

- Neural network algorithm for detection of new word meanings denoting named entities. *IEEE Access*, 10:68499–68512.
- Jader Camboim de Sá, Dimitra Anastasiou, Marcos Da Silveira, and Cédric Pruski. 2024. Socio-cultural adapted chatbots: Harnessing knowledge graphs and large language models for enhanced context awareness. In *Proceedings of the 1st Workshop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024)*, pages 21–27, St Julians, Malta. Association for Computational Linguistics.
- Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in characterization of semantic change. *ArXiv*.
- Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, Seng-Beng Ho, and San Murugesan. 2023. Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38:62–69.
- Lyle Campbell. 2013. *Historical linguistics*. Edinburgh University Press.
- Pierluigi Cassotti, Stefano De Pascale, and Nina Tahmasebi. 2024. Using synchronic definitions and semantic relations to classify semantic change types. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4539–4553, Bangkok, Thailand. Association for Computational Linguistics.
- Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *ArXiv*, abs/2311.09022.
- Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.
- Ishita Dasgupta, Andrew Kyle Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *ArXiv*, abs/2207.07051.
- Lauren Fonteyn and Enrique Manjavacas. 2021. Adjusting scope: A computational approach to case-driven research on semantic change. In *Workshop on Computational Humanities Research*.
- Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.
- Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. *ArXiv*, abs/2101.07668.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- Pride Kavumba, Ana Brassard, Benjamin Heinzerling, and Kentaro Inui. 2023. Prompting for explanations improves adversarial NLI. is this true? Yes it is true because it weakens superficial cues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2165–2180, Dubrovnik, Croatia. Association for Computational Linguistics.
- Kate Kearns. 2006. *The Handbook of English Linguistics*. Wiley Online Library.
- Adam Kilgarriff. 1997. I don’t believe in word senses. *Computers and the Humanities*, 31:91–113.
- George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.
- Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.
- Chenyang Lyu, Yongxin Zhou, and Tianbo Ji. 2022. MLLabs-LIG at TempoWiC 2022: A generative approach for examining temporal meaning shift. In *Proceedings of the First Workshop on Ever Evolving NLP (EvoNLP)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41.
- Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Cudas, Clarisse Simoes, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *ArXiv*, abs/2311.11045.

- Roberto Navigli. 2009. [Word sense disambiguation: A survey](#). *ACM Comput. Surv.*, 41:10:1–10:69.
- Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024a. [Analyzing semantic change through lexical replacements](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand. Association for Computational Linguistics.
- Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024b. [\(chat\)GPT v BERT dawn of justice for semantic change detection](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 420–436, St. Julian’s, Malta. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. [WiC: the word-in-context dataset for evaluating context-sensitive meaning representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.
- Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin UK.
- Lidia Pivovarovna and Andrey Kutuzov. 2021. [Rushifteval: a shared task on semantic shift detection for russian](#). *ArXiv*.
- Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. [Word sense disambiguation: A unified evaluation framework and empirical comparison](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [SemEval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.
- Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldböck, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte Im Walde. 2024. [The DUREL annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta. Association for Computational Linguistics.
- Gerard Steen, Lynne Cameron, Alan Cienki, Peter Crisp, Alice Deignan, Raymond W. Gibbs, Joe Grady, Zoltán Kövecses, Graham David Low, and Elena Semino. 2007. [Mip: A method for identifying metaphorically used words in discourse](#). *Metaphor and Symbol*, 22:1–39.
- James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S A Graziano, and Cristina Becchio. 2024. [Testing theory of mind in large language models and humans](#). *Nature human behaviour*.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. [Survey of computational approaches to lexical semantic change](#). *arXiv: Computation and Language*.
- Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2013. [Semantic change computation: A successive approach](#). *World Wide Web*, 19:375–415.
- Elizabeth Closs Traugott. 2017. [Semantic change](#).
- Ruiyu Wang and Matthew Choi. 2023. [Large language models on lexical semantic change detection: An evaluation](#). *ArXiv*, abs/2312.06002.
- Taylor W. Webb, Keith J. Holyoak, and Hongjing Lu. 2022. [Emergent analogical reasoning in large language models](#). *Nature Human Behaviour*, 7:1526–1541.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- Janyce Wiebe and Rada Mihalcea. 2006. [Word sense and subjectivity](#). In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia. Association for Computational Linguistics.
- Xi Ye and Greg Durrett. 2022. [The unreliability of explanations in few-shot prompting for textual reasoning](#). In *Neural Information Processing Systems*.
- Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. [Lscdiscovery: A shared task on semantic change discovery and detection in spanish](#). *ArXiv*, abs/2205.06691.
- Xulang Zhang, Rui Mao, Kai He, and Erik Cambria. 2023. [Neuro-symbolic sentiment analysis with dynamic word sense disambiguation](#). In *Conference on Empirical Methods in Natural Language Processing*.

A Negative Examples

In this section, we present some examples where the model failed to produce the correct form or to give an answer over the correctly produced reasoning.

A.1 Dimension

Below are some examples of dimension data:

Example #44 come

- **Come** out of the closet!
- He **came** singing down the road.

In sentence 1 it is used with the sense of emerge. — In sentence 2 it is used with the sense of move in a certain direction. — "He came out of the closet singing." The word 'come' doesn't make a bad pun, but it has different senses. It is still different.

A: different

In the example above, the zeugma was not correctly constructed, and it ignored the instruction, which only makes sense if it makes a bad pun.

Example #67 do

- All you ever **do** is surf the Internet.
- A 'B' grade **doesn't** suffice to get me into medical school.

In sentence 1 it is used with the sense of an action. — In sentence 2 is used with the sense of to suffice, to be enough. — "A' B' grade does not suffice for me to do what I want." The word 'do' makes a bad pun, so is different.

A: different

In the example above, it failed to build a zeugma using both contexts, even though the answer is correct.

A.2 Orientation

Below are some examples of orientation data.

Example #186 sympathy

- He does not resent their supervision or John's sometimes tiring **sympathy**.
- The group exhibits more concern over the possibility that the abolitionist might make a martyr of John than it does over the development of **sympathy** for him.

In the first sentence, 'sympathy' is used in a negative context, indicating annoyance or irritation. — In the second sentence, 'sympathy' is used to express understanding or concern. — Expressing understanding or concern (second sentence) is more positive than feeling annoyed or irritated.

A: positive

While the answer was correct, in the example above, the model didn't use antithesis for ordering senses.

B Inference Settings

Given hardware constraints, we used the gguf version of the models with llama.cpp⁵ library. In the Phi-3 (mini version), we used the 4k context with fp16 quantization. LLaMA-3-8b we used 8bit quantization and LLaMA-3-70b, 2bit quantization. All model weights were obtained from HuggingFace⁶.

We did all the experiments on a Tesla V100 with 32GB RAM. The inference for all the models and prompts took less than four days.

C Correlation Across Prompt Strategies and Models

In this section, we present the correlation plot between judgments across different models and different prompts.

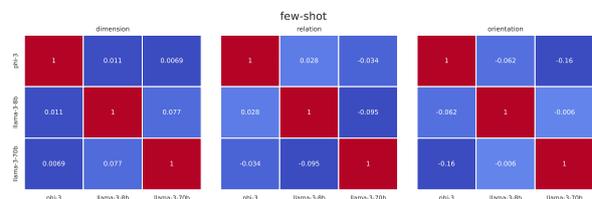


Figure 5: Correlation for Few-shot prompting between models across the three poles of change.

In Figure 10 we observe that few-shot and CoT approaches are highly correlated.

⁵<https://github.com/ggerganov/llama.cpp>

⁶<https://huggingface.co/models>

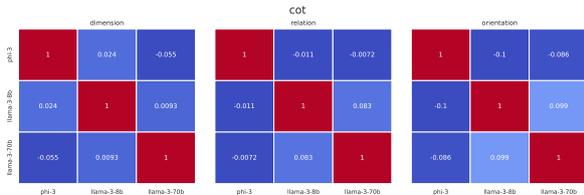


Figure 6: Correlation for CoT prompting across models.

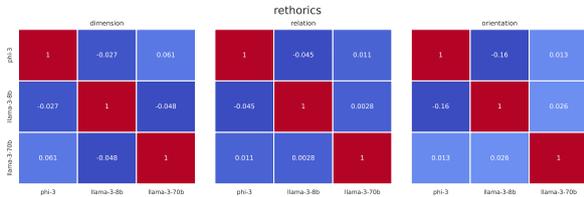


Figure 7: Correlation for Rhetorics prompting across models.

D Detailed Prompts

In the Figures 11 to 13, we show the detailed prompt for each type of characterization. For rhetorical prompt, traditional CoT prompt ("Let's think step-by-step") and direct prompt we provide the same few-shot examples.

E Dataset Creation and Annotation

The creation of the dimension dataset relies on the Word-in-Context data, however this data was originally built for sense relatedness with labels related and unrelated. In our problem settings we need identical or different under the zeugma analysis. So we agreed on annotating approximately 300 examples with about half as negative cases. For the identical cases the authors distributed randomly sentences until each author reached about 50 manually annotated examples pairs using the zeugma procedure described in (Kearns, 2006), some examples don't have enough context for easily using zeugma so they were discarded from the data. This procedure produced sentence pairs with "different" labels so the first author complete the rest with random examples from with the "unrelated" label, as they logically must be different if they are unrelated. Some cases were discarded when the author judged to have the same sense. This was performed until reaching about 300 examples in total. The authors have English as the second language (non-native speakers).

We created the relation data by using MASC and SEMCOR to craft examples.

To obtain sentiment labels for the orientation data the authors filtered sentences from online dic-

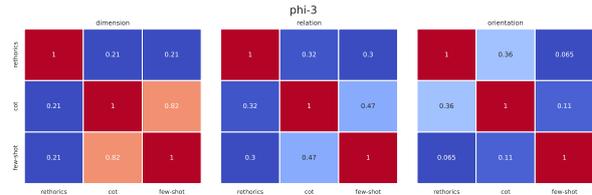


Figure 8: Correlation for Phi-3 model across prompt strategies.

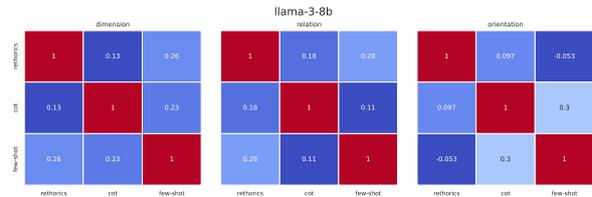


Figure 9: Correlation for LLaMA-3-8b model across prompt strategies.

tionaries with varying degrees of polarity.

We first provided the annotators with the agreement terms: "This is a study on sentiment perception of polysemous words. This data will be freely available for research purposes. Inside you'll be asked to rate how the feeling varies for a word in different sentences. Your answers will be completely anonymous. COMPANY will not collect your personal data through this questionnaire and will not be able to identify you based on your answers. For more information about COMPANY's privacy notice please visit our webpage at: URL"

Then we presented a training screen in Figure 14.

We prompted the annotators for sentiment analysis with screen Figure 15.

For each example a 3 annotation of sentiment was collected. With three annotators we observe Fleiss' Kappa of 0.3104 and Krippendorff of 0.3115.

F Ai Assistants In Research Or Writing

As our native language is not English, we used AI assistants like Grammarly, ChatGPT, and Gemini to improve vocabulary, grammar, and readability of this documents and prompts. We also checked all generated text for inconsistencies with the original intent and fixed them properly when identified.

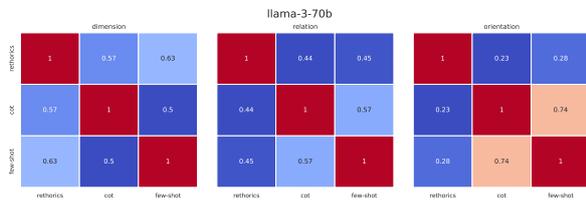


Figure 10: Correlation for LLaMA-3-70b model across strategies.

Sense Figurativeness

You are presented with two sentences that both contain a specific word. Your task is to analyze how this word is used in each sentence and determine if its usage in the second sentence represents a metaphor or a metonymy with respect to its use in the first sentence. Follow these steps to complete the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Compare the uses, determining if the second is related as a metaphor (where the word is used in a similar but non-literal sense), as a metonymy (where the word represents something closely related to or associated with it), or unrelated, used with a different sense.
- Step 4. Based on the third reasoning, write the final answer, 'metaphor', 'metonymy', or 'unrelated.'

[Few-shot examples.]

Figure 12: Prompt for figurative sense in the relation dataset.

Sense Differentiation

You are presented with two sentences that both contain a specific word. Your task is to analyze how this word is used in each sentence and determine if its usage in the second sentence represents the same sense with respect to its use in the first sentence. Follow these steps to complete the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Write a sentence that joins both sentences using zeugma and the same shared word while preserving the same sense. If the construction make a bad pun, the words have different sense.
- Step 4. Based on the previous reasoning give your final answer: 'identical' or 'different.'

[Few-shot examples.]

Figure 11: Prompt for sense differentiation in the dimension dataset.

Sense Orientation

You will be provided with two sentences that share a common word used with different senses. Your task is to describe if the second sense for the word is more positive than the first. Follow this instructions to execute the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Leverage the rhetorical strategy of antithesis, contrasting a negative with a positive, to weigh why one meaning might be more favorable than the other, or if they stand neutral.
- Step 4. Based on the third reasoning, write the final answer 'negative', 'positive', or 'neutral.'

[Few-shot examples.]

Figure 13: Prompt for sense orientation in the orientation dataset.

You are presented with two sentences that both contain a specific word. Your task is to analyze how this word is used in each sentence and determine if its usage in the second sentence is more positive, negative, or neutral with respect to the first sentence. Please consider only the feeling of the word in your judgement, not the whole sentence.

Example:

Sentence 1:

"As the hiker reached the summit, the breathtaking view of the majestic mountains left him in an awful silence, overwhelmed by the awe-inspiring beauty that surrounded him."

Sentence 2:

"Upon witnessing the gruesome battlefield, the medieval knight was struck with an awful realization of the carnage that lay before him."

The usage of 'awful' in the second sentence is more positive, negative or neutral with respect to the first sentence?

Answer:

In the first sentence the word 'awful' is used with the sense of full of admiration, while in the second sentence it is used with the meaning of horrible. So the **second** sentence is more negative.

Positive Negative Neutral Dumbo

Figure 14: Training screen

Please mark if positive, negative or neutral for the question below.

Word:

burn

Sentence 1:

The mutual aid is reciprocal and all have a burning desire to learn.

Sentence 2:

After visiting so many places today, my muscles are burning.

The usage of 'burn' in the second sentence is more positive, negative or neutral with respect to the first sentence?

Positive Negative Neutral I don't know

Figure 15: Annotation screen for sentiments.

Using BERT to Explore Lexical Semantic Change of Prepositions

Liudmila Radchankava¹ and Vasily Konovalov^{2,3}

¹Johannes Gutenberg University Mainz

²AIRI, ³MIRAI

luradcha@uni-mainz.de

Abstract

This paper presents a semi-supervised approach to explore the lexical semantic change of English prepositions using the BERT language model. We adopt a method that leverages BERT to obtain representations of prepositions, train a model on these representations, and use this model to measure the lexical semantic changes of the prepositions. We show that BERT embeddings can be used for preposition-sense disambiguation. Then, by applying the trained model to sentences extracted from the Corpus of Historical American English (COHA), we detect a variety of lexical semantic changes over time for English prepositions.

1 Introduction

Many linguists distinguish between lexical categories and functional categories of words. While functional categories have mostly grammatical meanings, lexical categories have a more obvious descriptive content. In the case of prepositions, linguists argue that there are two subclasses: function prepositions and content prepositions. Accordingly, some assume that the class of prepositions includes both grammatical and lexical elements, analogous to the distinction between content verbs and function verbs (auxiliaries) (Foucambert and Zuniga, 2012).

It has also been suggested that prepositions are a hybrid between the two categories depending on their usage. Therefore, the prepositions cluster can be located between grammatical words (like determiners) on the one hand, and lexical words (nouns, verbs, adjectives and adverbs) on the other. For example, the English preposition *of* is grammatical, whereas *off* is lexical (Boye and Harder, 2012).

It is broadly known that prepositions are a highly polysemic class (association of one word with two or more distinct meanings). The preposition *on*, for example, is originally a spatial adposition (*the book*

is on the table), it may express temporal relations (*it was on 10 March*), causal relations (*she read the book on my suggestion*), or relations of manner (*she spoke on a rising tone*) (Stefanowitsch et al., 2020).

Using the correct sense of a preposition is important for understanding the meaning of the utterance (Baldwin et al., 2009). This task becomes even more challenging because of the scarcity of the training dataset. Prepositions have not received adequate attention in the NLP community. Even in the Penn Treebank, only two types of prepositions are recognized (*in* (locative, temporal, and manner) and *to* (direction)) (O’Hara, 2005).

Bidirectional Encoder Representations from Transformers (BERT) are a language representation model. Unlike other models, BERT is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right contexts in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, including text classification, sequence classification, and question answering, without significant task-specific architecture modifications. Google AI released multiple versions of BERT, such as multilingual BERT, English BERT, and Chinese BERT (Google-Research, 2018; Devlin et al., 2019).

In this work we apply BERT to the task of preposition-sense disambiguation and measure lexico-semantic change of prepositions. Our contributions are:

1. We show that despite the mixed nature of prepositions, BERT-based contextualized word embeddings used by the kNN classifier can address preposition-sense disambiguation and achieve reasonable performance in comparison to the previous systems that leverage external resources. Also, we show that mul-

tilingual BERT performance falls behind the language-specific English base BERT by a small margin.

2. We propose a distance-to-labeled-exemplars signal derived from BERT embeddings as a proxy for diachronic contextual divergence in prepositions.
3. We provide empirical evidence that, under a distance-based proxy derived from contextual embeddings, compound prepositions show clearer diachronic shifts in usage contexts than simple prepositions, whose developments are typically more gradual and harder to detect automatically.

Even though our study was applied to English prepositions, the same methods can be applied to any language using multilingual BERT provided an appropriate preposition dataset.

2 Related Work

Exploring usage types of prepositions is a part of the Word Sense Disambiguation (WSD) task. The goal of WSD is to identify the correct sense of the word usage from a fixed set of sense identifiers. To train and evaluate WSD systems, many shared task datasets have been published in the SemEval workshop series.

Previous approaches to preposition-sense disambiguation (PSD) were based on external knowledge resources as WordNet (extracting synonyms and hypernyms) including extensive feature engineering. Instead of using BOW that focuses solely on the words within a fixed window size, they focus on elements that are related via the phrase structure (Tratz and Hovy, 2009).

Also, multilingual data can be used to improve the accuracy of PSD. The approach is to pretrain an LSTM encoder to predict the translation of a preposition, and then to incorporate the pre-trained encoder as a component in a supervised classification system, as well as fine-tune it for the task. The multilingual signals consistently improve the results on preposition-sense dataset (Gonen and Goldberg, 2016).

It has been shown that BERT embeddings can be utilized directly to approach the WSD task due to their property of providing distinct vector representations for the same token depending on its context. To learn the semantic capabilities of CWEs, a simple yet interpretable kNN approach has been

employed. However, the approach has only been tested on nouns (Wiedemann et al., 2019).

Moreover, BERT has been found useful in exploring over-time development of a preposition (only of *about*). The study focused on its three most frequent senses. It demonstrated that unsupervised BERT-embeddings can identify the usage types of *about* in English. It formulated three expected functional-semantic developments of *about* based on linguistic literature, and assessed whether BERT-embeddings can be used in combination with Entropy Difference measures to (semi-) automatically detect these changes in sentences between 1810 and 2009 taken from the Corpus of Historical American English (COHA) (Fonteyn, 2020).

The multilingual BERT was trained on Wikipedia for 104 languages and uses a shared cross-language BPE vocabulary. Its cross-lingual transferability was demonstrated on sequential tasks like NER and POS (Pires et al., 2019), and in multi-tasking mode (Karpov and Konovalov, 2023).

3 Datasets

To train the model we used The Preposition Project (TPP) dataset, which was the dataset for SemEval 2007 Task 6: Word-Sense Disambiguation of Prepositions. This corpus covers 34 prepositions with 16,557 training and 8,096 test sentences, each containing a single preposition example. The sentences were extracted from the FrameNet database, based mostly on the British National Corpus (designed to represent a wide cross-section of British English, both spoken and written, from the late twentieth century). Each preposition has a different set of possible senses. The prepositions were labeled in accordance with the Oxford Dictionary of English (ODE). We used the original split to train and test sets (Litkowski and Hargraves, 2007).

4 Experimental Setup

For PSD, we rely on contextualized representations of prepositions. We leverage the BERT language model to compute preposition representations for each occurrence in an utterance. Then we train the model given the contextualized preposition representation and corresponding semantic classes. Having trained the model we label all the given preposition use cases along the temporal axis. Then we detect and analyze the occurrences associated with larger distances to the nearest neighbor.

4.1 BERT Language Model

We produce usage representations using the BERT language model, a multi-layer bidirectional Transformer encoder trained on masked token prediction and next sentence prediction. We use the original English BERT base model `bert-base-uncased` with 12 layers, 768 hidden dimensions, and 110M parameters. To measure the effect of the number of parameters we test the large BERT model (`bert-large-uncased`). Also, we use the multilingual BERT model (`bert-base-multilingual-uncased`) to measure to what extent multilingual BERT can be applied to language-specific PSD.

Given a word of interest w and a context of occurrence $c = (v_1, \dots, v_n)$, with $w = v_i$, we extract the preposition embeddings from the last four hidden layers of the encoder and average them (Savkin et al., 2024). It appears that prepositions are frequent enough tokens and they are never divided into sub-tokens in our settings.

4.2 Nearest-Neighbor Classifier

We apply a non-parametric nearest-neighbor classifier (kNN) to label the preposition senses. The kNN classification algorithm (Cover and Hart, 1967) assigns a plurality vote of a sample’s nearest labeled neighbors. We select kNN classifier because its effectiveness for PSD is mainly based on the representation abilities of BERT. Moreover, it can directly lead to the training example that sets a certain classifier decision. In the most simple case, one-nearest neighbor, it predicts the label from the nearest training instance by some defined distance metric. As distance measure for kNN, we use the Euclidean distance of the CWE vectors.

Our approach considers only senses for a target word that has been observed during training. This approach has been named as localized nearest neighbor word sense disambiguation (Wiedemann et al., 2019).

4.3 Detecting Semantic Change

The TPP prepositions can be divided between simple and compound prepositions (from diachronic point of view). The simple category contains the most ancient prepositions, e.g. *in*, *on*, *of*, *by*, *with* etc. The compound prepositions, however, contains the majority of prepositions (cf. *about*, *across*, *along*, *around*, *aside*, *beside*, *onto*, *into*, *aback*, *above*, *after*, *again*, *apart*, *astray*, *asunder*,

athwart, *before*, *behind*, *below*, *between* etc.) (Lindstromberg, 2010). Simple prepositions could be viewed as grammatical items (they fulfill the valency of a verb e.g. to belong to smb, to depend on smth/smb, to fear for smb), while compound prepositions are still between two poles and can have both grammatical and lexical characteristics (Lehmann, 2015). It can be assumed that simple prepositions tend to exhibit more gradual, highly polysemous, and context-dependent semantic developments, which are harder to capture using distributional methods, whereas compound prepositions are more likely to undergo discretely identifiable lexico-semantic changes (Lehmann, 2015; Rhee, 2002). This does not imply that simple prepositions do not undergo semantic change. Rather, their changes are often incremental and organized in radial semantic structures, making them less salient for automatic detection based on contextual embedding distances. We analyze five representative prepositions from each group (*in*, *on*, *at*, *with*, *of*) and five representative compound prepositions (*into*, *onto*, *beside*, *above*, *about*). Accordingly, cases where OED-documented senses predate the benchmark period do not contradict our findings, as the proposed proxy captures shifts in usage contexts rather than the emergence of entirely new dictionary senses.

For each preposition, we sample 2,000 sentences from the Corpus of Historical American English (COHA). The corpus is balanced by genre across decades; sentences were sampled uniformly from COHA without additional temporal stratification. COHA is the largest structured corpus of historical English. It contains more than 475 million words of text from the 1820s-2010s (which makes it 50-100 times as large as other comparable historical corpora of English) and the corpus is balanced by genre decade by decade (Davies, 2012).

We apply BERT to extract CWE of prepositions from the sampled sentences, then we measure the distance between the preposition representation from the sample and its top-1 neighbor from TPP. Larger distances correspond to lower classifier confidence, which we interpret as a proxy for contextual divergence and potential diachronic change.

As the last step, we compute Spearman’s rank correlation between the usage year and the distance to the top-1 neighbor. Spearman’s ρ tests for a monotonic association without assuming normality. We report two-sided p -values and treat correlations with $p \leq 0.05$ as statistically significant.

5 Experimental Results

We conduct two types of experiments. We calculate the performance of the kNN model on the SemEval dataset to justify that our settings can be applied to solve PSD despite mixed lexico-grammatical origins of the prepositions. In our second experiment, we use the trained model to label sampled prepositions from COHA over time. Furthermore, we identify the old-fashioned usages by leveraging the classifier’s confidence (the distance to the nearest neighbor).

5.1 Nearest-Neighbor Model

In the SemEval corpus, each preposition has a different set of senses, and the natural approach is to learn a different model for each preposition.

To obtain a more robust nearest-neighbor classification, we optimize for the k parameter. The optimization is performed on the development set (80%/20% split of the original training set).

Table 1 shows results of k optimization for the English base BERT model, large BERT model, multilingual BERT. The best result, with an accuracy of 80.2, is achieved for the base BERT model and $k = 5$. Surprisingly, the base BERT model outperforms the large BERT model by a small margin, whereas the large BERT model has a comparable performance with the multilingual BERT model for $k > 6$.

Table 2 compares results of the base BERT model on the test set with those of previous systems. The First Sense baseline selects the first sense of each preposition as the answer (under the assumption that the senses are organized somewhat according to prominence). The Freq Sense baseline selects the most frequent sense from the training set.

The KU system uses a statistical language model based on a large unannotated corpus. The model is used to evaluate the likelihood of various substitutes for a word in a given context. These likelihoods are then used to determine the best sense for the word in novel contexts (Yuret, 2007).

The MELB-YB system achieved the highest result out of the three participating systems in the SemEval 2007 shared task. The approach leveraged features such as POS tags and WordNet-based features, and also high-level features (semantic role tags), using a word window of up to seven words, in a Maximum Entropy classifier (Ye and Baldwin, 2007).

Model	<i>base-BERT</i>	<i>large-BERT</i>	<i>M-BERT</i>
k=1	79.4	76.9	73.9
k=2	77.3	75.3	72.3
k=3	79.7	77.5	75.4
k=4	79.6	77.5	75.5
k=5	80.2	77.4	75.5
k=6	80	77.2	75.4
k=7	80	77.1	76
k=8	79.6	76.9	75.4
k=9	79.5	76.8	75.3
k=10	79.5	76.7	75.2

Table 1: Optimizing for k parameter.

Model	Accuracy
Our system (bert-base-uncased)	79.8
LSTM+ (Gonen and Goldberg, 2016)	81.7
MELB-YB (Ye and Baldwin, 2007)	69.3
KU (Yuret, 2007)	54.7
IRST-BP (Popescu et al., 2007)	49.6
Most Frequent Sense	39.6
First Sense	28.9

Table 2: The accuracy on the test set of the TPP dataset (the SemEval corpus).

The approach that pretrains an LSTM encoder to predict the translation of a preposition, incorporating the pre-trained encoder as a component in a supervised classification system and fine-tuning it for the task, achieves 81.7 (Gonen and Goldberg, 2016).

Our kNN base BERT model achieves an accuracy of 79.8. It outperforms all SemEval 2007 participants, though it falls short of the LSTM-based model that leverages external multilingual data. In contrast, our models do not include any external knowledge, based solely on CWE of BERT.

The goal of this comparison is not to achieve the state-of-the-art performance but to show that BERT CWE along with the kNN classifier can be used to solve PSD, which is different from noun or verb sense disambiguation due to the mixed lexico-grammatical roles of prepositions in sentences.

5.2 Detecting Semantic Change

In our second experiment, we label all the utterance samples from COHA with the corresponding preposition classifier and assign a distance to the nearest

neighbor (a measure of confidence). Then we calculate the correlation between these distances and the usage year taken from COHA. Table 3 shows the Spearman correlation coefficient with the corresponding p -value. According to the correlation results we have two groups. The group of simple prepositions (*on*, *at*, *in*, *of*, *with*) either shows low correlation close to 0, or has a high p -value, which means that null hypothesis is correct (no correlation). This pattern is consistent with the idea that, under our distance-based proxy, simple prepositions show weaker or less consistently detectable diachronic trends, while their changes may be more gradual and harder to capture automatically (Lehmann, 2015; Rhee, 2002). However, all compound prepositions except for *into* have a negative correlation, suggesting that earlier attestations tend to be associated with larger distances (lower confidence) under our proxy.

Preposition	Correlation
<i>on</i>	0.06($3.8 \cdot 10^{-3}$)
<i>in</i>	-0.005(NS)
<i>at</i>	-0.06($2.4 \cdot 10^{-3}$)
<i>of</i>	0.05(NS)
<i>with</i>	0.05($4.1 \cdot 10^{-2}$)
<i>into</i>	0.075($4.91 \cdot 10^{-12}$)
<i>above</i>	-0.13($5.24 \cdot 10^{-37}$)
<i>onto</i>	-0.10($3.19 \cdot 10^{-11}$)
<i>beside</i>	-0.22($4.56 \cdot 10^{-43}$)
<i>about</i>	-0.13($1.18 \cdot 10^{-35}$)

Table 3: Spearman correlation coefficient with corresponding p -value; NS indicates $p > 0.05$.

onto The preposition *onto* is used with verbs to express movement on or to a particular place or position: *Move the books onto the second shelf*. The preposition *onto* is also used to show that something faces in a particular direction: *The window looked out onto the terrace*. Larger distances to the nearest neighbor for *onto* were mainly observed in sociolectal (non-standard register) contexts.

I wuz a settin onto a rockincheer, and Hanner Ann wuz on my knee. (1865)

... the stoppin to pick up them ez coodent stick onto ther flyin steeds, I hed no difficulty in outrunnin em. (1866)

... vagary uv the mind, wich, wen loosed from its clay, sores off onto its own hook, without any

restraint. Is the giant Republican actually dead. (1869)

above According to OED the preposition *above* expresses position in or movement to a place that is higher: *The water came above our knees*; more than something; greater in number, level or age than somebody/something: *Inflation is above 6%*. Larger distances to the nearest neighbor were caused by a slightly outdated preposition usage: *the splintered rapiers of their opponents appeared a speedy omen of the enforcement of the threats of utter annihilation, that was hoarsely shrieked above the tumult of the contest*. Here *above* is used as in examples "shout above the noise" or rather "shout over the noise". The next interesting usage, "not being above four feet" for "being less than four feet": *south shore was much more elevated than that near the river, which had here extremely low banks, the water in the stream not being above four feet below them*

beside Preposition *beside* is used in a sense of next to or at the side of somebody/something: *He sat beside her all night*; compared with somebody/something: *My painting looks childish beside yours*. Larger distances to the nearest neighbor were often observed in contexts with a nearby pronoun, which may reflect contextual similarity patterns in the TPP benchmark rather than the emergence of a new sense.

scene was going on in the cell of the destined victim. His daughter kneeled beside him at daylight in his prison. She had cheered his solitude with the sunshine. (1832)

sister Bel has reason to be thankful, " said Catharine, who was close beside her sister, " for your teaching her name so familiarly to the river-gods. (1831)

into According to OED the preposition *into* expresses motion from without to a point within limits of space, time, condition, circumstance, etc.; the motion which results in the position expressed by *in*, or which is directed towards that position: *Come into the house.*; in the direction of something: *Speak clearly into the microphone*. Preposition *into* occurs most frequently among compound prepositions and seems to be better integrated into the grammatical system. Accordingly, *into* shows a comparatively weak trend under our proxy (despite statistical significance), suggesting a relatively stable usage pattern. However, there are some out-

dated examples:

[...]in nations; and he was convinced that, before vice could be thus exalted into custom, there must exist in the community which would tolerate such an institution (1816)

It is possible that these circumstances may have occasionally betrayed me into intemperances of expression which I did not intend: it is certain, that I (1817)

5.2.1 Limitations and Interpretation

The distance to the nearest labeled neighbor should not be interpreted as a direct measure of semantic innovation. Instead, it serves as a proxy for contextual divergence from a fixed, contemporary sense inventory derived from the SemEval-2007 dataset. Importantly, the SemEval-2007 sense inventory is substantially imbalanced: frequent senses are more densely represented in embedding space and therefore tend to yield smaller nearest-neighbor distances. For this reason, we interpret distances comparatively over time within the same preposition, rather than as absolute indicators of novelty. Moreover, increasing distances may reflect not only semantic reanalysis but also contextual expansion, stylistic drift, register change, or corpus composition effects. Consequently, the proposed method is intended as an exploratory tool for identifying diachronic shifts in usage patterns rather than a definitive test of lexical semantic change in the lexicographic sense.

6 Conclusion

Some of the detected effects, particularly for prepositions such as *beside*, may reflect contextual expansion rather than the emergence of entirely new senses as documented in historical dictionaries such as the OED. Our distance-based measure is sensitive to shifts in usage contexts, which can manifest as increasing semantic conventionalization or stylistic change. In this sense, our results capture a broader notion of language change, encompassing both semantic reanalysis and contextual expansion. In this work, we measured to what extent BERT-based contextualized word embeddings can address preposition-sense disambiguation. To test their capabilities to distinguish different preposition senses we used a kNN classifier. For our experiments, we used TPP datasets from SemEval 2007 task 6. We can conclude that the kNN model trained on BERT embeddings achieves reasonable

performance, despite the mixed lexico-grammatical origins of prepositions. Also, by using the trained model we observe a lexical semantic change in compound prepositions and analyze it ad hoc.

References

- Timothy Baldwin, Valia Kordoni, and Aline Villavicencio. 2009. [Prepositions in applications: A survey and introduction to the special issue](#). *Comput. Linguistics*, 35(2):119–149.
- Kasper Boye and Peter Harder. 2012. [A usage-based theory of grammatical status and grammaticalization](#). *Language*, 88:1–44.
- Thomas M. Cover and Peter E. Hart. 1967. [Nearest neighbor pattern classification](#). *IEEE Trans. Inf. Theory*, 13(1):21–27.
- Mark Davies. 2012. [The 400 million word corpus of historical american english \(1810–2009\)](#). pages 231–262.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Lauren Fonteyn. 2020. [What about grammar? using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions](#). In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*, volume 2723 of *CEUR Workshop Proceedings*, pages 257–268. CEUR-WS.org.
- Denis Foucambert and Michael Zuniga. 2012. [Effects of grammatical categories on letter detection in continuous text](#). *Journal of psycholinguistic research*, 41:33–49.
- Hila Gonen and Yoav Goldberg. 2016. [Semi supervised preposition-sense disambiguation using multilingual data](#). In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2718–2729. ACL.
- Google-Research. 2018. [Tensorflow code and pre-trained models for bert](#). <https://github.com/google-research/bert>.
- Dmitry Karpov and Vasily Konovalov. 2023. [Knowledge transfer between tasks and languages in the multi-task encoder-agnostic transformer-based models](#). In *Computational Linguistics and Intellectual Technologies*, volume 2023.

- Christian Lehmann. 2015. *Thoughts on grammaticalization*. Number 1 in Classics in Linguistics. Language Science Press, Berlin.
- Seth Lindstromberg. 2010. [English prepositions explained: Revised edition](#). pages 1–273.
- Kenneth C. Litkowski and Orin Hargraves. 2007. [Semeval-2007 task 06: Word-sense disambiguation of prepositions](#). In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23-24, 2007*, pages 24–29. The Association for Computer Linguistics.
- Thomas P O’Hara. 2005. *Empirical acquisition of conceptual distinctions via dictionary definitions*. Ph.D. thesis, Citeseer.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual bert?](#) In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4996–5001. Association for Computational Linguistics.
- Octavian Popescu, Sara Tonelli, and Emanuele Pianta. 2007. [IRST-BP: preposition disambiguation based on chain clarifying relationships contexts](#). In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23-24, 2007*, pages 191–194. The Association for Computer Linguistics.
- Seongha Rhee. 2002. Semantic changes of english preposition against: A grammaticalization perspective.
- Maksim Savkin, Anastasia Voznyuk, Fedor Ignatov, Anna Korzanova, Dmitry Karpov, Alexander Popov, and Vasily Konovalov. 2024. [DeepPavlov 1.0: Your gateway to advanced NLP models backed by transformers and transfer learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 465–474, Miami, Florida, USA. Association for Computational Linguistics.
- Anatol Stefanowitsch, Elena Smirnova, and Matthias Hüning. 2020. *2 Complex adpositions in three West Germanic Languages: German, Dutch, and English*, pages 65–138. De Gruyter Mouton, Berlin, Boston.
- Stephen Tratz and Dirk Hovy. 2009. [Disambiguation of preposition sense using linguistically motivated features](#). In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA, Student Research Workshop and Doctoral Consortium*, pages 96–100. The Association for Computational Linguistics.
- Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. [Does BERT make any sense?](#) [interpretable word sense disambiguation with contextualized embeddings](#). In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.
- Patrick Ye and Timothy Baldwin. 2007. [MELB-YB: preposition sense disambiguation using rich semantic features](#). In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23-24, 2007*, pages 241–244. The Association for Computer Linguistics.
- Deniz Yuret. 2007. [KU: word sense disambiguation by substitution](#). In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23-24, 2007*, pages 207–214. The Association for Computer Linguistics.

A Computational Analysis of the Emergence of Therapy-speak in Social Media

Alina Iacob

Ana Sabina Uban

University of Bucharest, Faculty of Mathematics and Computer Science,

HLT Research Center

alinaiacob9517@gmail.com, auban@fmi.unibuc.ro

Abstract

The present article comparatively investigates semantic change in psychology-related concepts across scientific and social media texts. We assess patterns of change over 15 years and compare word usage corpora of psychology journals abstracts and Reddit comments, testing whether specialized communities on social media align with psychology experts. Our results confirm that many terms are used differently on social media compared to specialized literature. Furthermore, Reddit communities focused on psychology discussions occupy an intermediate position, adopting a more objective stance than general-domain threads while remaining distinct from specialized literature.

1 Introduction and Related Work

In the current century, mental health has steadily become a main concern. As the population is confronted with what is deemed a "mental-health crisis" (Evans et al., 2018; Schnittker, 2025), considerable efforts have been made to increase awareness regarding psychological issues (Foulkes and Andrews, 2023). On the one hand, this has led to undeniably positive outcomes, such as an increase in seeking professional help (Foulkes and Andrews, 2023). On the other hand, some scholars claim that the increased presence of the mental health-related discourse in the everyday life can lead to negative consequences such as people's excessively interpreting their normal negative emotions as mental disorders (Foulkes and Andrews, 2023), or the trivialization of the severity of the conditions (Spencer and Carel, 2021). Another consequence of the permeation of medical terminology into the mainstream is the advent of *therapy speak* - "the imprecise and superficial integration of psychotherapy language into everyday communication" (Isern-Mas and Almagro, 2025). While semantic change, understood as "a change in the concepts associated with a word" (Lehmann, 2013),

is a naturally-occurring linguistic phenomenon, the situation in which a term is simultaneously used with a different connotation by experts and laymen can raise concerns - as even medical experts report difficulties in drawing the boundary between disorder and normal manifestations of negative emotions (Wakefield and First, 2013). Regarding psychology-related words in particular, the "concept creep" hypothesis posits that harm-related concepts are particularly prone to broaden their meaning over time so as to encompass both new and milder phenomena (Haslam, 2016). Subsequent research has shown that some concepts do indeed undergo such a transformation in specialized literature (Haslam, 2016).

In previous research, Vylomova and Haslam (2021) studied the concept creep hypothesis in relation to five harm-related concepts via Word2Vec embeddings trained on two different corpora, and found that concepts show a tendency towards becoming wider and milder, confirming the concept creep hypothesis. The same tendency was observed by Baes et al. (2023b). Baes et al. (2023a) evaluated the phenomenon of pathologisation, by which traits generally considered normal human behaviours become conceptualised as mental disorders. Pisl et al. (2025) analyzed the semantic severity of anxiety, depression, and trauma on a corpus of New York Times. A study conducted on Reddit data (Record et al., 2018) found that the online discourse of self-identified counseling experts aligns with established counseling principles.

Since no previous study has compared the use of psychology-related terminology in specialized literature with data from social media, our study aims to bridge this gap using an up-to-date corpus. Specifically, we aim to:

- 1) Analyze whether psychology-related terms undergo similar patterns of semantic transformation over time in both professional literature and social media, testing the concept creep hypothesis.

2) Determine whether, within the same time period, the use of psychology-related terms differs significantly between the specialized corpus and the general domain.

Drawing from the methodology employed by [Vylomova and Haslam \(2021\)](#), we analyze concepts previously identified as prone to misuse or change ([Isern-Mas and Almagro, 2025](#); [American Psychological Association, 2024](#)): *toxic, bipolar, psychopath, narcissistic, triggered, trauma, OCD, imposter, gaslighting, addiction, bullying, harassment, and prejudice*. While previous research has examined the general domain by turning to corpora based on magazines or newspapers, we aim to answer the aforementioned questions by analyzing the proposed terms in social media. To assess whether specialized communities within the Reddit platform utilize the proposed terms in a similar manner with psychology specialists, we create two separate Reddit corpora by making a distinction between general threads and threads dedicated to psychology discussions. Moreover, in addition to previous work, we adapt the methodology to use contextual embeddings alongside static ones.

2 Data Collection and Preprocessing

Psychology Corpus (PC). For the specialized scientific corpus, we start from the dataset compiled by [Vylomova et al. \(2019\)](#), which spans the period from 1930 to 2019 and includes psychology abstracts sourced from the E-Research and PubMed databases, and we further expand the corpus by extracting additional abstracts from PubMed up to 2025. To align with the availability of social media data, we restrict the corpus to abstracts published from 2010 onward.

Reddit General Corpus (RGC) and Reddit Psychology Corpus (RPC). For the social media data analysis, we assemble a corpus of Reddit comments. To investigate the presence of specialized communities within Reddit, we divide the data into two separate corpora according to subreddits: the Reddit General Corpus (RGC) and the Reddit Psychology Corpus (RPC), including only psychology-oriented subreddits. For all corpora, we exclude entries shorter than 15 words. Descriptive statistics are presented in the Appendix (Figs. 3-5).

For all experiments, the target concepts are mapped to a single canonical form. A full description of the mapping can be found in the Appendix, Table 2. For the experiments based on static em-

beddings, the following preprocessing steps were employed: removing stopwords, punctuation and numeric tokens, then lemmatizing using SpaCy.

3 Methodology

We follow the methodology established by [Vylomova and Haslam \(2021\)](#) to evaluate four dimensions of semantic change: relative term frequency, semantic breadth, semantic displacement, and the evolution of neighbor similarity. We use the same methodology described here to train Word2Vec models on our corpora, and we additionally employ contextual embeddings based on pretrained BERT models ([Devlin et al., 2019](#)). For a selection of terms, we compare generic pretrained models with domain-specific models pretrained on psychology data (MentalBERT ([Ji et al., 2022](#))).

Global Word2Vec Model. We train a global type-level Word2Vec model on each corpus.

Year-Specific Word2Vec Models. We train year-specific Word2Vec models for each year, initializing from the global model and fine-tuning on sentences from the respective year.

Contextual Embeddings. For each target word and year, we: (i) filter sentences containing the word; (ii) randomly sample 10% of sentences; (iii) extract embeddings for the word (averaging over subtokens); (iv) compute the mean embedding for the sample; (v) repeat sampling 5 times; (vi) aggregate all sample-level embeddings by averaging across samples.

3.1 Semantic Breadth

Semantic breadth refers to the range of contexts each target concept is used in. According to the concept creep hypothesis, health-related concepts tend to widen their breadth over time ([Haslam, 2016](#)).

We capture the semantic breadth by calculating the cosine similarity between the contexts each concept is used in. A low similarity points to the fact that the concept is used across different contexts (wide semantic breadth), while a high similarity shows that the concept is confined to specific contexts (narrow semantic breadth). We obtain context representations as follows:

Static Embeddings. For each concept and year, up to 20 sentences containing the target word are sampled. A 5-token window around the target word (tokens present in the year-specific Word2Vec vocabulary) is converted to a vector by averaging

token embeddings. This sampling and aggregation were repeated 10 times, producing sub-aggregate vectors per repetition. Semantic breadth is quantified as the mean pairwise cosine similarity of all sub-aggregate vectors per year.

Contextual Embeddings. For each concept and year, sentences containing the target word are identified, and a 10% random sample (minimum 1 sentence) are taken. A 5-token window around the target word is extracted from each sentence. Word or context embeddings are obtained from BERT, averaging subword embeddings per sentence. The sampling is repeated 5 times, and all embeddings are averaged to produce a single mean embedding per concept/year. Semantic breadth is quantified as the mean pairwise cosine similarity among all sentence embeddings for that year.

3.2 Semantic Displacement

Semantic displacement quantifies how much each concept has changed from one year to another, by computing the cosine distance between corresponding word representations in each year.

3.3 Neighbours Similarity Evolution

We aim to capture how the relationship between the concept and words it is often associated with changes over time in each corpus. We first construct a list of top 5 nearest neighbours for each concept in each corpus, and compute the evolution of the cosine similarity between the concept and each neighbour over time. We rely solely on static embeddings for these experiments, since they are more suitable for retrieving global nearest neighbors (relative to the entire vocabulary) efficiently.

4 Results

4.1 Frequency Analysis

While each concept behaves differently, an overall upward trend in relative frequency was observed for the proposed terms. The terms *bipolar*, *trauma*, and *bullying* showed increasing frequencies across all three corpora. Notably, *trauma* is the most frequent term in the PC data, whereas in the RPC, it only becomes the most frequent term starting with 2016. The term *toxic* has a particularly sharp increase in the RGC, becoming the most frequent among the proposed terms, while in the RPC it has a slower increase and remained at all times surpassed in frequency by other terms. In the PC, *toxic* maintains its frequency. The frequency of the

term *narcissistic* shows a decreasing tendency in the PC, while its presence increases in the RPC and RGC corpora. Exact frequencies plots are found in the Appendix. All frequency plots are smoothed with a 3-year centered moving average.

Table 1: Semantic Breadth: Trend and Volatility

Concept	PC	RGC	RPC
Addiction	Broadening (Stable)	Narrowing (Stable)	Narrowing (Stable)
Bipolar	Broadening (Stable)	Narrowing (Mod. Volatile)	Narrowing (Stable)
Bullying	Broadening (Mod. Volatile)	Narrowing (Stable)	Narrowing (Mod. Volatile)
Gaslighting	Broadening (High Volatile)	Broadening (High Volatile)	Broadening (Mod. Volatile)
Harassment	Narrowing (Mod. Volatile)	Narrowing (Stable)	Narrowing (Mod. Volatile)
Imposter	Broadening (High Volatile)	Narrowing (High Volatile)	Broadening (High Volatile)
Narcissistic	Broadening (Mod. Volatile)	Narrowing (Stable)	Narrowing (Stable)
OCD	Narrowing (Mod. Volatile)	Narrowing (Stable)	Narrowing (Stable)
Prejudice	Narrowing (Stable)	Narrowing (High Volatile)	Narrowing (Mod. Volatile)
Psychopath	Broadening (Stable)	Broadening (Mod. Volatile)	Broadening (Stable)
Toxic	Narrowing (Mod. Volatile)	Narrowing (Stable)	Broadening (Mod. Volatile)
Trauma	Broadening (Stable)	Broadening (Stable)	Narrowing (Stable)
Triggered	Broadening (Stable)	Narrowing (Stable)	Narrowing (Mod. Volatile)

4.2 Semantic Breadth

Since contextual embeddings uncover certain patterns more clearly than static embeddings, particularly in the displacement experiments, we focus this analysis on contextual embeddings. Contextual embeddings have also become the standard method for semantic representations in lexical semantic change studies in recent years (Periti and Tahmasebi, 2024). Results obtained using static models are reported in the Appendix (Figures 9-14), alongside results obtained with MentalBERT (Figures 18-20). We note that there is no substantial difference between the results obtained using the BERT-base model and those derived from MentalBERT, indicating that the base model is sufficient for capturing semantic shifts in psychology-related vocabulary. All semantic breadth plots are smoothed with a 3-year centered moving average. Figure 1 shows the evolution of semantic breadth for each concept. To quantify shifts in the corpora, we applied linear regression with year as the independent variable and mean cosine similarity as the dependent variable. For each concept, the regression slope captures the annual change in semantic breadth, reflecting whether the concept’s contextual usage is expanding or contracting over time. To measure the overall magnitude of variation, we computed the amplitude as the difference between the maximum

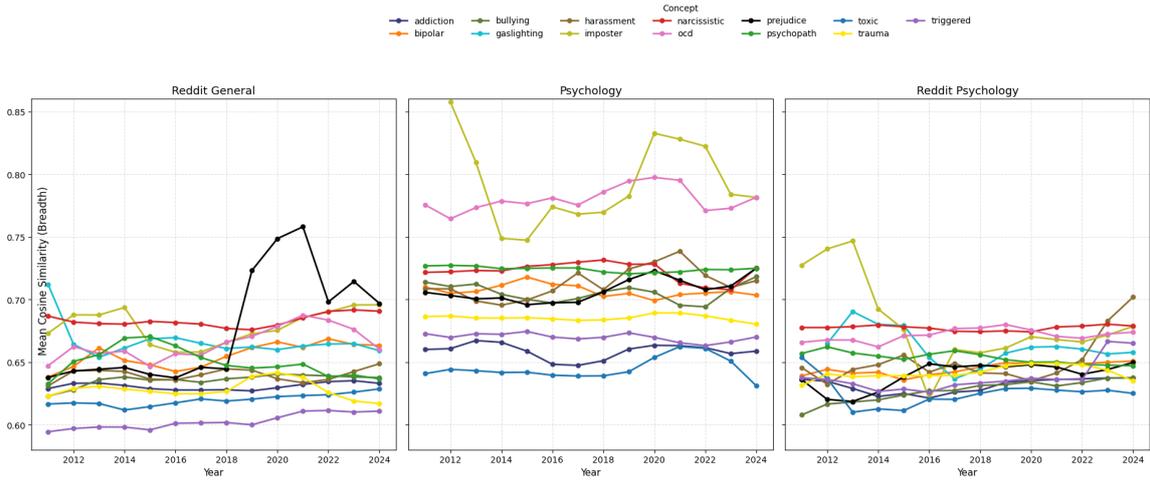


Figure 1: Semantic Breadth using BERT

and minimum similarity values observed for each concept across the available years. Table 1 summarizes the results based on this analysis for each term. The exact values for the regression slope and amplitude are found in the Appendix (Table 3).

In the Reddit Corpora, there is a higher number of terms presenting a narrowing in breadth, which mirrors the results obtained by Vylomova and Haslam (2021). For instance, the terms *triggered* and *narcissistic* widen their breadth in the PC, but their breadth becomes narrower in the RGC. This inverse relationship suggests that specialized terms may be adopted into general use while being applied in only a subset of their original contexts.

As in the case of the terms' frequencies, the RPC occupies an intermediary position between the specialized PC and the RGC: some of the terms follow the patterns present in the specialized corpus, while others follow the general domain trends.

4.3 Semantic Displacement

Figure 3 presents the semantic displacement results obtained with the contextual embeddings. The results obtained with Word2Vec year models can be found in the Appendix (Figures 15-17).

The specialized PC is the steadiest in terms of displacement. There is little change in long-established psychological terms such as *OCD*, *bipolar*, or *trauma*, but terms such as *gaslighting* or *imposter* shift their semantics substantially from one year to another, possibly pointing to the fact that they still haven't sedimented into a fixed usage.

Terms in the RGC show the most displacement, while the RPC occupies once more an intermediary position. We also note an inverse relationship between increases in term frequency and semantic displacement in the Reddit corpora. For instance, the term *toxic* has an abrupt increase in frequency in the RGC around the year 2016, which is the same year when the semantic displacement begins to stabilize. This suggests that starting with 2016 the term acquired a dominant meaning that is predominantly used thereafter.

4.4 Neighbours Similarity Evolution

To understand the evolution of neighbours in the embeddings space for each concept, we illustrate the evolution of a selection of nearest neighbors (Appendix, Figs. 22-33). One example of the trajectory of the most relevant neighbors for such a

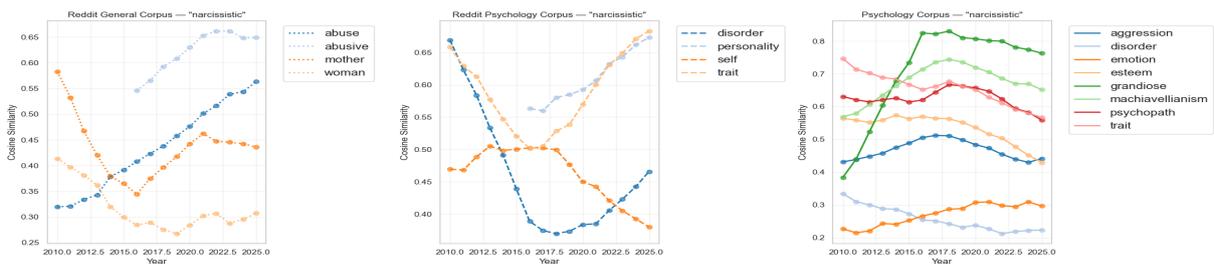


Figure 2: Narcissistic - Neighbours Similarity Evolution

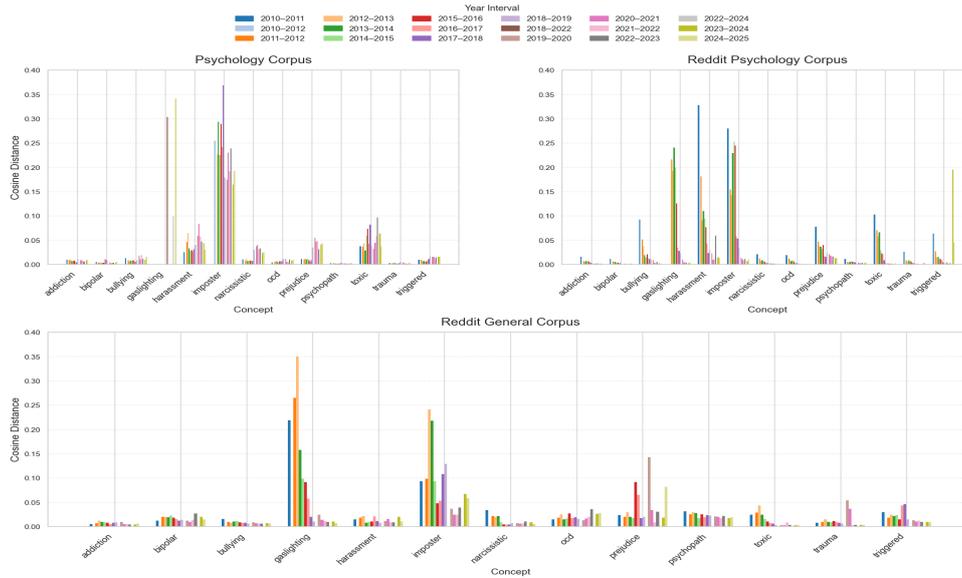


Figure 3: Semantic Displacement using BERT

term is shown in Figure 2 for "narcissistic". The complete list with the top neighbours for each year can also be found in the Appendix, Tables 4-42.

While some words show similar associations across all three corpora (e.g., *bipolar*, *psychopath*), others are used in markedly different contexts. For example, in the PC, the term *toxic* is not linked to psychological traits or behaviors, but to harmful substances. In contrast, in the RGC, its usage shifts noticeably: around 2010, it is associated with terms such as "chemical" and "health", but starting in 2015, there is a sharp increase in associations with words like "masculinity" and "relationship" - coinciding with the year when *toxic* experiences a rapid rise in frequency in the Reddit Corpus. The RPC mirrors this trend, following the pattern observed in the RGC.

Another term with a contrastive usage is *narcissistic* (Fig. 3). While in the RGC it is associated with words such as "abuse" and "mother", RPC and PC present a more objective stance, the term being associated with words such as "disorder" or "trait".

5 Conclusions

Our experiments provide insights into the nature and mechanisms of semantic change. The results show that concepts shift semantically both diachronically throughout years and concurrently - when they are adopted from a specialized domain into general usage.

There is not one single pattern for semantic change. Consistent with previous findings, we

have found that the concept creep phenomenon is more prevalent in specialized literature than in general usage. We have also shown that semantic displacement experiments can point to the concepts that are not traditionally related to psychology vocabulary, as these terms tend to present a greater degree of displacement. Our results indicate that some terms adopted from specialized terminology are integrated into everyday language with distinct nuances (such as *narcissistic*, *addiction*, etc.) In contrast, other words—although appearing in psychology-related contexts, as evidenced by their presence in the RPC, were likely not derived from specialized psychology vocabulary, either because their usage in the specialized corpus differs substantially (as in the case of *toxic*) or is entirely absent (as with *gaslighting*).

Finally, we find that there is a difference in word usage between general Reddit comments and comments extracted from forums dedicated to Psychology, the latter having a somewhat intermediary position between specialized usage and general domain usage. Thus, our research brings insights into the interaction between cultural shifts and semantic change. Based on these results, research in the fields of psychology, sociology and linguistics can further determine whether the divergence in word usage between specialized literature and general discourse reflects misuse - as in the so-called *therapy speak* - or represents a natural process of cultural change.

Limitations

Due to computational limitations, we restricted the context selected for extracting contextual embeddings to a sample of the original corpus for each term in the case of some analyses, which might affect the robustness of the results. Similarly, we extract contextual embeddings as well as contexts for computing semantic breadth using static embeddings based on a 5-word context window. Especially for contextual embeddings, the limited context could affect the quality of the resulting representations - using entire sentences as contexts instead could be a future improvement.

One of the dataset limitations is focusing solely on abstracts extracted from Psychology journals, which might not reflect the full range of terms' usage. Moreover, the identified differences between Reddit corpora and scientific abstracts could reflect a mix of the assumed effect of the authors' expertise and competence in using psychological terms (experts versus laypeople) and the general differences in communicative function and genre convention in the two types of texts. While we do find that there are differences in term usage even within Reddit (by comparing RG and RPG), future work could further examine this, possibly by selecting Reddit authors based on verified expertise (Record et al., 2018).

Ethical Statement

All data used in the study is anonymous. No personally identifiable information was accessed, stored, or processed during the research. Access to the dataset can be granted upon request for academic and research purposes only.

Acknowledgements

We would like to thank the reviewers for their insightful and detailed comments and suggestions. This research was partly supported by the Ministry of Education and Research, CNCS- UEFISCDI, project SIROLA, number PN-IV-P1- PCE-2023-1701, within PNCDI IV, and by the UEFISCDI Project "Centru de Excelență pentru Schimbări Climatice și Sociale (CESCS)", project number PN-IV-P6-6.1-CoEx-2024-0042.

References

- American Psychological Association. 2024. [How to harness the power of therapy-speak](#). *Monitor on Psychology*, 55(6). Accessed: 2025-12-20.
- Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2023a. Semantic shifts in mental health-related concepts. In *Proceedings of the 4th workshop on computational approaches to historical language change*, pages 119–128.
- Naomi Baes, Ekaterina Vylomova, Michael Zyphur, and Nick Haslam. 2023b. The semantic inflation of “trauma” in psychology. *Psychology of Language and Communication*, 27(1):23–45.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Teresa M Evans, Lindsay Bira, Jazmin Beltran Gastelum, L Todd Weiss, and Nathan L Vanderford. 2018. Evidence for a mental health crisis in graduate education. *Nature biotechnology*, 36(3):282–284.
- Lucy Foulkes and Jack L Andrews. 2023. Are mental health awareness efforts contributing to the rise in reported mental health problems? a call to test the prevalence inflation hypothesis. *New Ideas in Psychology*, 69:101010.
- Nick Haslam. 2016. Concept creep: Psychology’s expanding concepts of harm and pathology. *Psychological inquiry*, 27(1):1–17.
- Carme Isern-Mas and Manuel Almagro. 2025. Unmasking therapy-speak. *Theoretical Medicine and Bioethics*, 46(6):465–489.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [Mental-BERT: Publicly available pretrained language models for mental healthcare](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.
- Winfred P Lehmann. 2013. *Historical linguistics: An introduction*. Routledge.
- Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282.
- Vojtech Pisl, Ana-Maria Bucur, and Ioana R Podina. 2025. Revisiting the semantic severity of anxiety and depression: Computational linguistic study of normalization and pathologization. *Journal of Medical Internet Research*, 27:e73950.

Rachael A Record, Will R Silberman, Joshua E Santiago, and Taewook Ham. 2018. I sought it, i reddit: Examining health information engagement behaviors among reddit users. *Journal of health communication*, 23(5):470–476.

Jason Schnittker. 2025. The sociology of mental health and the twenty-first-century mental health crisis. *Society and Mental Health*, 15(1):1–16.

Lucienne J Spencer and Havi Hannah Carel. 2021. ‘isn’t everyone a little ocd?’: the epistemic harms of wrongful depathologisation. *Philosophy of Medicine*, 2(1).

Ekaterina Vylomova and Nick Haslam. 2021. Semantic changes in harm-related concepts in english. *Computational approaches to semantic change*, 6:93.

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st international workshop on computational approaches to historical language change*, pages 29–34.

Jerome C Wakefield and Michael B First. 2013. Clarifying the boundary between normality and disorder: a fundamental conceptual challenge for psychiatry.

A Appendix

Models and configurations used were as follows:

- BERT models: google-bert/bert-base-uncased¹ and mental/mental-bert-base-uncased². We use the final layer of the models as the embedding representation (as preliminary experiments showed it is more useful than middle layers for our analyses).
- Word2vec hyperparameters: window size 10, minimum count 5, 10 epochs.

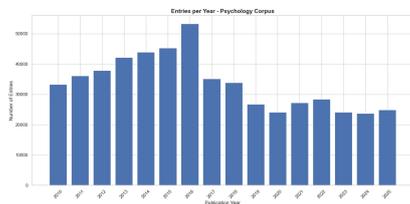


Figure 4: Entries per Year - Psychology Corpus

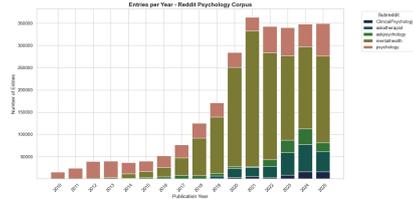


Figure 5: Entries per Year - Reddit Psychology Corpus

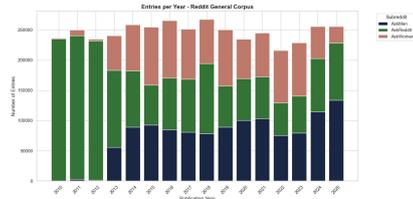


Figure 6: Entries per Year - Reddit General Corpus

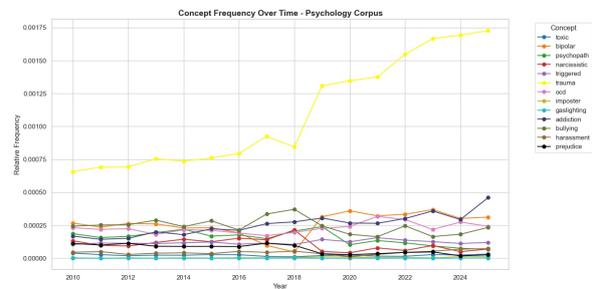


Figure 7: Frequency Analysis - Psychology Corpus

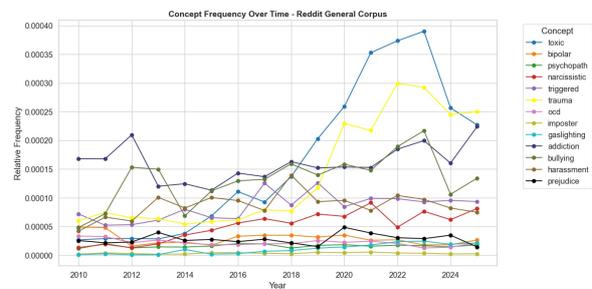


Figure 8: Frequency Analysis - Reddit General Corpus

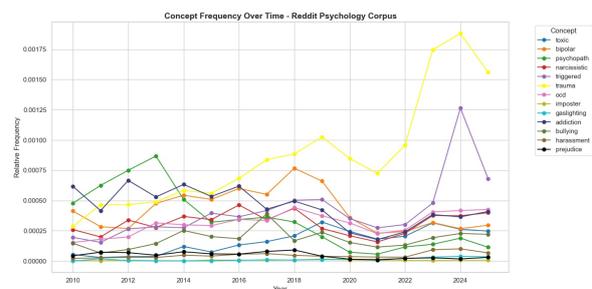


Figure 9: Frequency Analysis - Reddit Psychology Corpus

¹<https://huggingface.co/google-bert/bert-base-uncased>

²<https://huggingface.co/mental/mental-bert-base-uncased>

Concept	Mapped Forms
Psychopath	psychopath, psychopaths, psychopathy, psychopathic
Narcissistic	narcissist, narcissists, narcissistic, narcissism
Bipolar	bipolar, bipolarity, bipolar disorder, bipolar disorders
Toxic	toxic, toxicity
Triggered	trigger, triggers, triggered, triggering
Trauma	trauma, traumas, traumatic, traumatized, traumatised, post-traumatic, post-trauma, traumatize, traumatizes
OCD	obsessive compulsive disorder, obsessive compulsive disorders, obsessive compulsive, OCD
Imposter	imposter, impostor, imposters, impostors, imposter syndrome, impostor syndrome, imposter syndromes, impostor syndromes, imposter phenomenon, impostor phenomenon
Gaslighting	gaslighting, gaslighted, gaslights, gaslight
Addiction	addict, addiction, addicts
Bullying	bully, bullying, bullies
Harassment	harass, harassment, harasses
Prejudice	prejudice

Table 2: Target concepts and their corresponding lexical variants

Table 3: Quantitative Metrics: Slope and Amplitude across Corpora

Concept	Slope			Amplitude		
	PC	RGC	RPG	PC	RGP	RPG
Addiction	-0.0001	0.0003	0.0006	0.037	0.014	0.028
Bipolar	-0.0007	0.0019	0.0010	0.027	0.070	0.026
Bullying	-0.0001	0.0009	0.0021	0.055	0.030	0.050
Gaslighting	-0.0198	-0.0034	-0.0011	0.207	0.144	0.111
Harassment	0.0013	0.0009	0.0031	0.103	0.046	0.112
Imposter	-0.0018	0.0009	-0.0043	0.223	0.089	0.310
Narcissistic	-0.0001	0.0003	0.0001	0.052	0.018	0.016
OCD	0.0010	0.0018	0.0008	0.073	0.072	0.034
Prejudice	0.0016	0.0059	0.0015	0.045	0.279	0.057
Psychopath	-0.0003	-0.0002	-0.0008	0.010	0.082	0.021
Toxic	0.0000	0.0011	-0.0013	0.065	0.025	0.080
Trauma	-0.0002	-0.0001	0.0006	0.015	0.045	0.036
Triggered	-0.0003	0.0013	0.0018	0.018	0.025	0.116

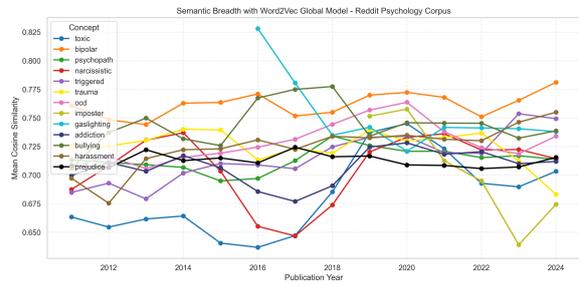


Figure 12: Semantic Breadth with Word2Vec Global Model - Reddit Psychology Corpus

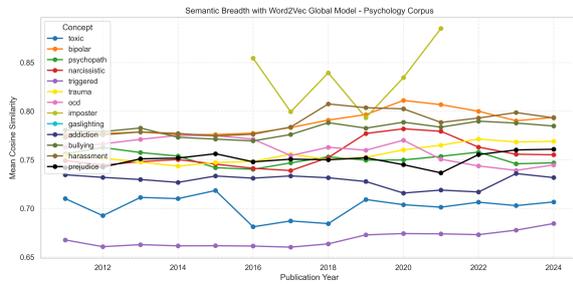


Figure 10: Semantic Breadth with Word2Vec Global Model - Psychology Corpus

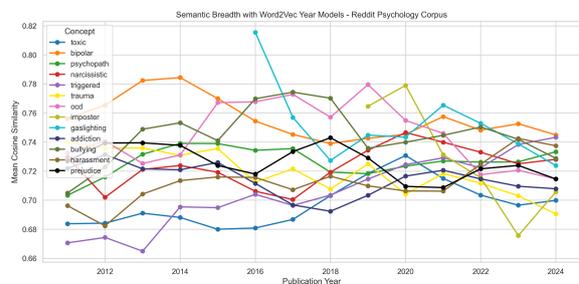


Figure 13: Semantic Breadth with Word2Vec Year Models - Reddit Psychology Corpus

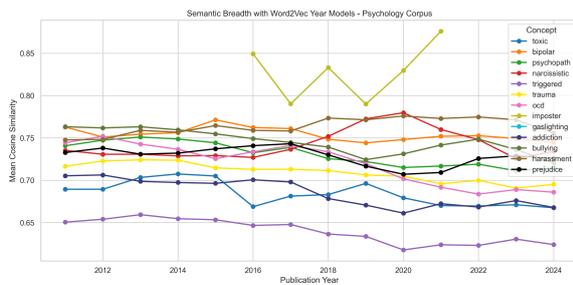


Figure 11: Semantic Breadth with Word2Vec Year Models - Psychology Corpus

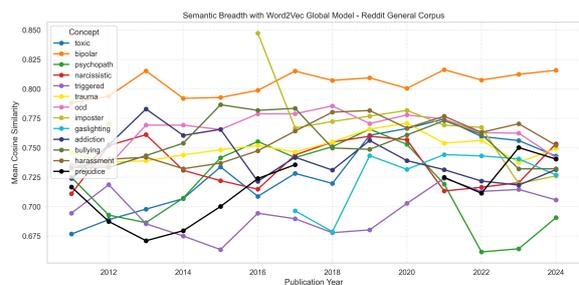


Figure 14: Semantic Breadth with Word2Vec Global Model - Reddit General Corpus

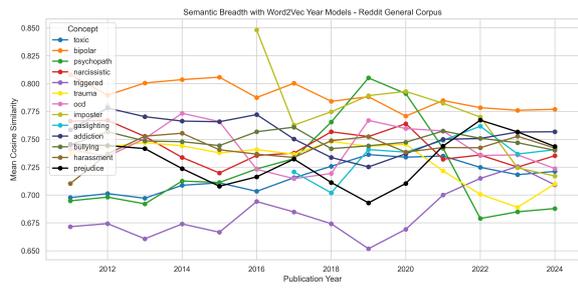


Figure 15: Semantic Breadth with Word2Vec Year Models - Reddit General Corpus

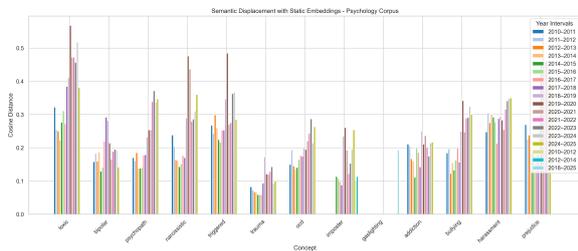


Figure 16: Semantic Displacement with Word2Vec Year Models - Psychology Corpus

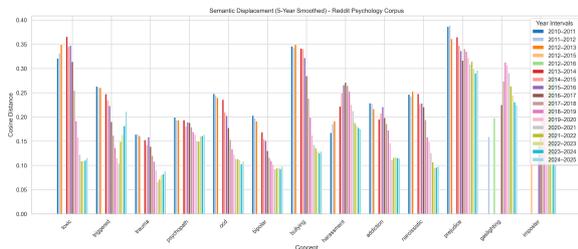


Figure 17: Semantic Displacement with Word2Vec Year Models - Reddit Psychology Corpus

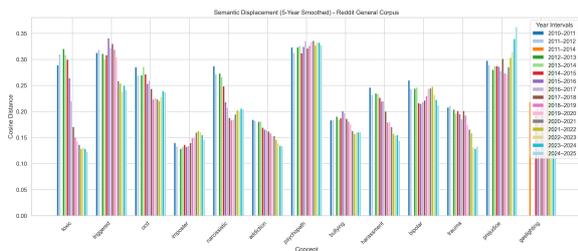


Figure 18: Semantic Displacement with Word2Vec Year Models - Reddit General Corpus

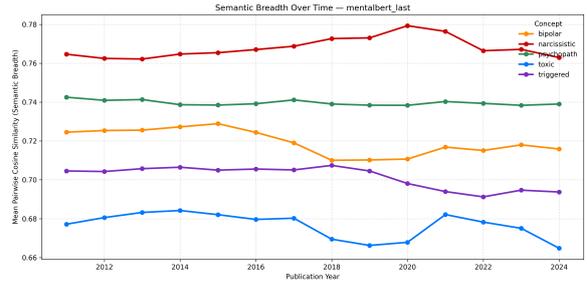


Figure 19: Semantic Breadth with MentalBERT - Psychology Corpus

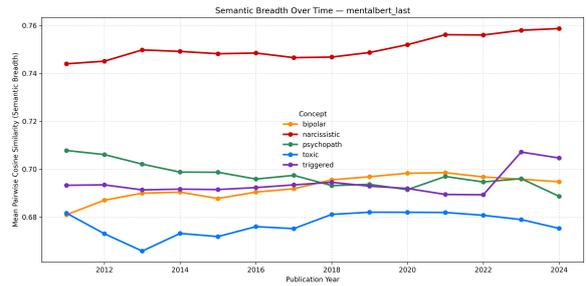


Figure 20: Semantic Breadth with MentalBERT - Reddit Psychology Corpus

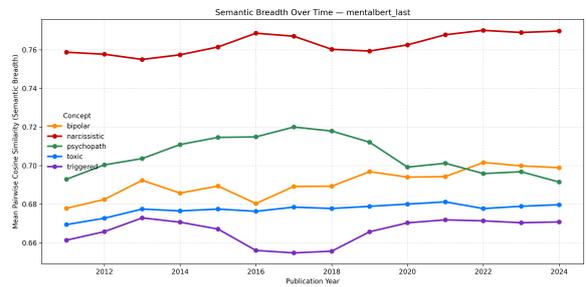


Figure 21: Semantic Breadth with MentalBERT - Reddit General Corpus

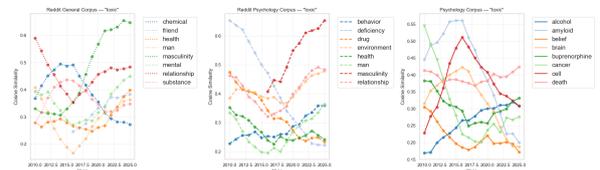


Figure 22: Toxic - Neighbours Similarity Evolution

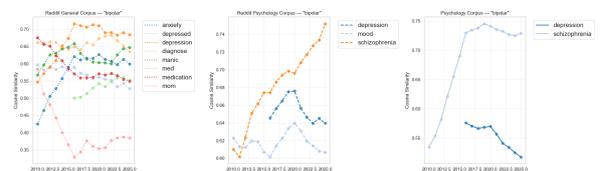


Figure 23: Bipolar - Neighbours Similarity Evolution

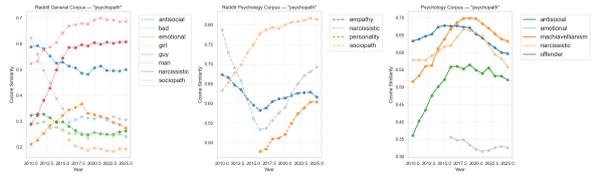


Figure 24: Psychopath - Neighbours Similarity Evolution



Figure 25: Triggered - Neighbours Similarity Evolution

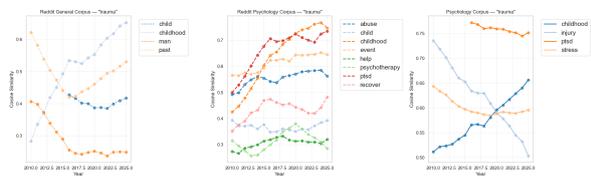


Figure 26: Trauma - Neighbours Similarity Evolution

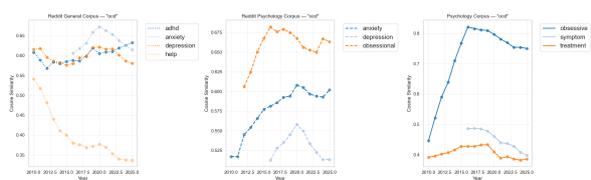


Figure 27: OCD - Neighbours Similarity Evolution

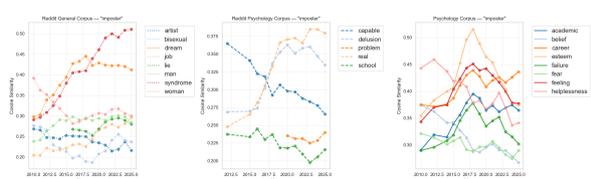


Figure 28: Imposter - Neighbours Similarity Evolution

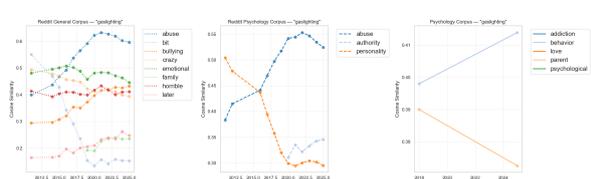


Figure 29: Gaslighting - Neighbours Similarity Evolution

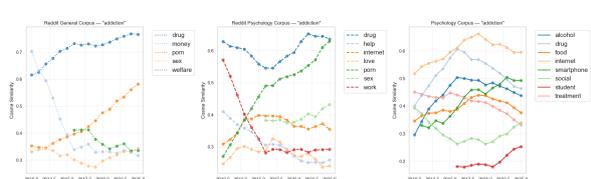


Figure 30: Addiction - Neighbours Similarity Evolution



Figure 31: Bullying - Neighbours Similarity Evolution

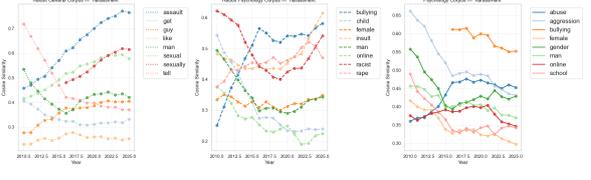


Figure 32: Harassment - Neighbours Similarity Evolution

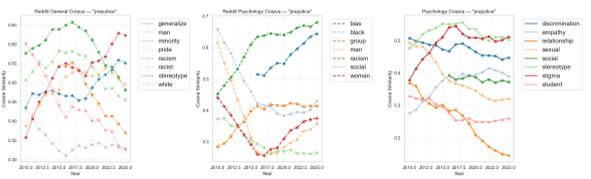


Figure 33: Prejudice - Neighbours Similarity Evolution

Table 4: Top-5 Co-occurring Neighbors per Year for *addiction* - Psychology Corpus

Year	1	2	3	4	5
2010	treatment	drug	study	disorder	internet
2011	treatment	drug	internet	use	study
2012	internet	treatment	drug	study	disorder
2013	drug	internet	study	treatment	disorder
2014	internet	study	use	treatment	drug
2015	treatment	study	internet	use	drug
2016	study	internet	treatment	food	drug
2017	internet	food	treatment	use	study
2018	internet	food	study	use	social
2019	study	internet	treatment	use	disorder
2020	treatment	study	use	disorder	drug
2021	study	use	disorder	treatment	alcohol
2022	study	internet	treatment	use	smartphone
2023	study	internet	smartphone	social	disorder
2024	internet	social	study	disorder	scale
2025	internet	study	social	scale	student

Table 5: Top-5 Co-occurring Neighbors per Year for *bipolar* - Psychology Corpus

Year	1	2	3	4	5
2010	disorder	patient	study	control	depression
2011	disorder	patient	study	depression	control
2012	disorder	patient	study	depression	risk
2013	disorder	patient	study	depression	control
2014	disorder	patient	study	depression	depressive
2015	disorder	patient	study	depression	depressive
2016	disorder	patient	study	depression	schizophrenia
2017	disorder	patient	study	schizophrenia	depression
2018	disorder	patient	study	depression	schizophrenia
2019	disorder	patient	schizophrenia	study	depression
2020	disorder	schizophrenia	patient	study	depression
2021	disorder	patient	schizophrenia	study	depression
2022	disorder	patient	schizophrenia	study	depression
2023	disorder	schizophrenia	study	patient	depression
2024	disorder	schizophrenia	study	patient	depressive
2025	disorder	schizophrenia	study	patient	depressive

Table 6: Top-5 Co-occurring Neighbors per Year for *bullying* - Psychology Corpus

Year	1	2	3	4	5
2010	school	victim	study	victimization	child
2011	school	study	victim	report	victimization
2012	school	study	victim	student	child
2013	school	victim	victimization	child	study
2014	school	victim	victimization	study	student
2015	school	victimization	victim	study	child
2016	victimization	school	study	victim	student
2017	school	study	victimization	student	victim
2018	school	victimization	study	victim	behavior
2019	behavior	victimization	study	relationship	school
2020	victimization	school	study	student	high
2021	victimization	behavior	study	report	school
2022	victimization	school	study	child	behavior
2023	victimization	study	school	adolescent	high
2024	victimization	school	study	social	adolescent
2025	school	study	victimization	high	experience

Table 7: Top-5 Co-occurring Neighbors per Year for *gaslighting* - Psychology Corpus

Year	1	2	3	4	5
2010					
2011					
2012					
2013					
2014					
2015					
2016					
2017					
2018	clinical	paper	parent	child	clinician
2019					
2020					
2021					
2022	behavior	result	indicate	clear	sequential
2023					
2024	specifically	respondent	experience	frequent	psychological
2025	love	addiction	gaslighte	perceive	acceptability

Table 8: Top-5 Co-occurring Neighbors per Year for *harassment* - Psychology Corpus

Year	1	2	3	4	5
2010	sexual	study	workplace	experience	woman
2011	sexual	gender	experience	man	study
2012	sexual	experience	student	school	violence
2013	sexual	gender	study	woman	behavior
2014	sexual	experience	report	study	gender
2015	sexual	bullying	study	online	victim
2016	sexual	experience	study	behavior	report
2017	sexual	study	woman	experience	risk
2018	sexual	experience	woman	study	report
2019	sexual	female	study	victimization	aggression
2020	sexual	violence	experience	base	analysis
2021	sexual	experience	study	student	abuse
2022	sexual	violence	report	study	experience
2023	sexual	experience	violence	victimization	study
2024	sexual	violence	experience	student	abuse
2025	sexual	health	experience	bullying	study

Table 9: Top-5 Co-occurring Neighbors per Year for *imposter* - Psychology Corpus

Year	1	2	3	4	5
2010	feeling	examine	patient	objective	study
2011					
2012	phenomenon	self	university	research	efficacy
2013	belief	replace	capgras	syndrome	characterize
2014	tendency	person	phenomenon	job	satisfaction
2015	academic	self	concept	path	phenomenon
2016	phenomenon	career	self	feeling	fear
2017	self	esteem	phenomenon	level	instability
2018	phenomenon	self	student	esteem	feeling
2019	phenomenon	scale	fraudulence	systematic	examine
2020	phenomenon	individual	feeling	syndrome	association
2021	phenomenon	mindset	failure	fear	goal
2022	phenomenon	learn	helplessness	model	experience
2023	study	instrument	age	reliably	detect
2024	phenomenon	psychological	career	decision	resilience
2025	phenomenon	self	career	satisfaction	esteem

Table 10: Top-5 Co-occurring Neighbors per Year for *narcissistic* - Psychology Corpus

Year	1	2	3	4	5
2010	self	personality	study	high	esteem
2011	self	personality	study	high	examine
2012	personality	self	study	trait	disorder
2013	personality	self	study	high	associate
2014	self	personality	study	trait	psychopath
2015	self	study	trait	psychopath	personality
2016	self	study	personality	trait	high
2017	self	personality	psychopath	study	trait
2018	self	study	grandiose	vulnerable	personality
2019	psychopath	personality	machiavellianism	self	trait
2020	trait	personality	regulation	psychopath	emotion
2021	self	study	trait	personality	psychopath
2022	personality	study	trait	grandiose	self
2023	disorder	trait	personality	self	grandiose
2024	study	trait	personality	machiavellianism	psychopath
2025	trait	personality	study	aggression	high

Table 11: Top-5 Co-occurring Neighbors per Year for *ocd* - Psychology Corpus

Year	1	2	3	4	5
2010	disorder	obsessive	compulsive	symptom	patient
2011	disorder	symptom	obsessive	compulsive	study
2012	disorder	obsessive	compulsive	patient	symptom
2013	disorder	obsessive	patient	compulsive	symptom
2014	disorder	symptom	compulsive	obsessive	patient
2015	disorder	patient	symptom	obsessive	compulsive
2016	disorder	compulsive	obsessive	symptom	patient
2017	disorder	symptom	patient	compulsive	obsessive
2018	disorder	symptom	obsessive	compulsive	study
2019	patient	disorder	symptom	obsessive	compulsive
2020	disorder	patient	compulsive	obsessive	symptom
2021	disorder	patient	symptom	obsessive	compulsive
2022	disorder	patient	compulsive	obsessive	symptom
2023	disorder	patient	compulsive	obsessive	symptom
2024	disorder	patient	compulsive	obsessive	study
2025	disorder	symptom	obsessive	treatment	compulsive

Table 15: Top-5 Co-occurring Neighbors per Year for *trauma* - Psychology Corpus

Year	1	2	3	4	5
2010	symptom	ptsd	stress	study	injury
2011	ptsd	stress	symptom	disorder	study
2012	ptsd	symptom	stress	study	disorder
2013	ptsd	stress	symptom	study	disorder
2014	ptsd	stress	disorder	symptom	study
2015	symptom	stress	ptsd	disorder	experience
2016	symptom	stress	ptsd	study	disorder
2017	symptom	experience	study	stress	ptsd
2018	symptom	stress	experience	study	ptsd
2019	ptsd	stress	disorder	symptom	childhood
2020	ptsd	stress	symptom	disorder	study
2021	ptsd	symptom	stress	disorder	exposure
2022	ptsd	symptom	stress	disorder	study
2023	ptsd	symptom	stress	disorder	childhood
2024	symptom	stress	ptsd	disorder	study
2025	stress	ptsd	symptom	disorder	experience

Table 12: Top-5 Co-occurring Neighbors per Year for *prejudice* - Psychology Corpus

Year	1	2	3	4	5
2010	study	participant	group	social	high
2011	study	sexual	relationship	social	contact
2012	study	social	group	contact	sexual
2013	study	relationship	social	group	reduce
2014	study	social	group	intergroup	sexual
2015	study	group	social	contact	attitude
2016	study	group	sexual	social	research
2017	social	study	contact	group	effect
2018	study	social	group	contact	people
2019	social	tom	study	level	sensitivity
2020	laïcité	social	study	outcome	associate
2021	empathy	study	social	patient	intergroup
2022	people	self	study	discrimination	individual
2023	discrimination	stereotype	study	experience	group
2024	social	discrimination	continuum	distance	study
2025	scale	study	discrimination	stigma	student

Table 16: Top-5 Co-occurring Neighbors per Year for *triggered* - Psychology Corpus

Year	1	2	3	4	5
2010	response	suggest	stimulus	event	task
2011	process	suggest	response	stimulus	study
2012	process	suggest	effect	result	response
2013	process	response	stimulus	behavior	study
2014	study	process	stimulus	response	task
2015	response	suggest	study	result	stimulus
2016	response	effect	result	study	process
2017	study	response	result	social	process
2018	response	social	experience	suggest	result
2019	asmr	response	stimulus	study	cue
2020	response	study	symptom	stimulus	result
2021	response	stress	study	behavior	result
2022	response	stress	change	social	study
2023	self	effect	behavior	study	response
2024	social	response	symptom	lead	result
2025	study	effect	response	stress	anxiety

Table 13: Top-5 Co-occurring Neighbors per Year for *psychopath* - Psychology Corpus

Year	1	2	3	4	5
2010	study	personality	trait	measure	high
2011	trait	personality	study	high	factor
2012	trait	personality	study	associate	measure
2013	trait	study	personality	high	factor
2014	trait	personality	study	measure	high
2015	trait	study	personality	measure	narcissistic
2016	trait	study	personality	measure	high
2017	trait	narcissistic	study	machiavellianism	high
2018	trait	study	personality	narcissistic	machiavellianism
2019	trait	study	antisocial	behavior	individual
2020	trait	study	relate	high	emotional
2021	trait	study	personality	factor	model
2022	trait	offender	study	high	research
2023	trait	study	high	personality	report
2024	trait	study	machiavellianism	personality	narcissistic
2025	trait	study	association	relate	associate

Table 17: Top-5 Co-occurring Neighbors per Year for *addiction* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	drug	self	know	thing	work
2011	drug	like	people	disorder	work
2012	drug	like	people	think	sex
2013	people	drug	like	think	internet
2014	drug	people	like	help	mental
2015	drug	people	brain	like	mental
2016	drug	people	like	think	disease
2017	book	people	drug	like	love
2018	drug	people	like	think	thing
2019	drug	people	like	think	help
2020	drug	like	people	think	help
2021	drug	like	people	help	think
2022	drug	like	people	help	thing
2023	like	people	drug	porn	thing
2024	like	people	drug	porn	thing
2025	like	people	drug	porn	feel

Table 14: Top-5 Co-occurring Neighbors per Year for *toxic* - Psychology Corpus

Year	1	2	3	4	5
2010	cell	abeta	effect	brain	amyloid
2011	effect	cell	result	study	induce
2012	effect	amyloid-	brain	study	drug
2013	effect	oligomer	environment	suggest	study
2014	effect	cell	tau	stress	formation
2015	effect	study	cell	increase	child
2016	effect	exposure	health	study	child
2017	stress	pride	treatment	risk	white
2018	stress	art	alcohol	belief	interactive
2019	alcohol	stress	death	matter	opioid
2020	patient	response	effect	cancer	expectancy
2021	case	relate	buprenorphine	trauma	death
2022	stress	program	study	youth	associate
2023	financial	phenibut	cancer	study	day
2024	drug	health	case	death	mental
2025	drug	case	health	associate	exposure

Table 18: Top-5 Co-occurring Neighbors per Year for *bipolar* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	disorder	depression	mood	people	schizophrenia
2011	disorder	depression	diagnose	people	schizophrenia
2012	disorder	people	depression	diagnose	schizophrenia
2013	disorder	depression	people	schizophrenia	like
2014	disorder	depression	like	people	diagnose
2015	disorder	depression	like	people	diagnose
2016	disorder	depression	like	people	diagnose
2017	disorder	depression	like	people	diagnose
2018	disorder	depression	like	diagnose	people
2019	disorder	depression	like	diagnose	people
2020	disorder	depression	like	diagnose	people
2021	disorder	depression	like	diagnose	people
2022	disorder	depression	like	people	diagnose
2023	disorder	depression	like	people	diagnose
2024	disorder	like	depression	diagnose	people
2025	disorder	like	bpd	diagnose	depression

Table 19: Top-5 Co-occurring Neighbors per Year for *bullying* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	inner	think	school	need	self
2011	people	feel	think	way	like
2012	school	people	like	think	kid
2013	school	kid	like	people	child
2014	school	kid	people	like	think
2015	people	school	think	high	self
2016	people	school	like	think	feel
2017	and/or	life	requirement	conflict	pick
2018	people	school	like	kid	feel
2019	school	people	like	get	think
2020	school	people	like	think	get
2021	people	like	school	feel	get
2022	people	school	like	get	think
2023	people	school	like	get	feel
2024	school	people	like	get	feel
2025	people	school	like	get	feel

Table 23: Top-5 Co-occurring Neighbors per Year for *narcissistic* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	personality	disorder	people	self	trait
2011	self	high	personality	people	disorder
2012	self	personality	people	disorder	think
2013	self	personality	people	disorder	like
2014	people	personality	disorder	self	like
2015	people	self	like	disorder	think
2016	people	personality	think	self	disorder
2017	people	personality	disorder	self	like
2018	people	self	personality	like	disorder
2019	people	personality	like	disorder	self
2020	people	like	think	personality	disorder
2021	people	like	think	personality	disorder
2022	people	like	disorder	think	personality
2023	people	like	personality	think	self
2024	people	like	personality	disorder	think
2025	people	like	personality	trait	disorder

Table 20: Top-5 Co-occurring Neighbors per Year for *gaslighting* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010					
2011	idea	try	gaslighte	person	think
2012	following	warning	sign	true	dance
2013	wiki	prescriptive	authority	rxp(http://en.wikipedia.org)	wikipedia
2014	classic	turn	will	able	tell
2015	thing	try	understand	attempt	somebody
2016	abuse	mean	thing	think	wrong
2017	like	think	need	people	personality
2018	think	try	abuse	need	feel
2019	try	like	know	abuse	time
2020	try	people	like	feel	think
2021	people	like	feel	think	know
2022	like	people	try	think	know
2023	like	think	people	feel	try
2024	try	people	like	think	feel
2025	people	try	think	like	feel

Table 24: Top-5 Co-occurring Neighbors per Year for *ocd* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	disorder	people	anxiety	know	behavior
2011	anxiety	like	disorder	think	thing
2012	like	anxiety	people	thought	disorder
2013	pure	disorder	obsessional	like	people
2014	anxiety	disorder	like	depression	thought
2015	like	disorder	anxiety	thought	think
2016	anxiety	like	disorder	thought	people
2017	anxiety	like	thought	disorder	think
2018	anxiety	like	disorder	thought	people
2019	anxiety	like	thought	disorder	think
2020	like	anxiety	thought	disorder	think
2021	thought	like	anxiety	people	depression
2022	thought	like	anxiety	people	disorder
2023	like	thought	anxiety	people	think
2024	like	anxiety	thought	think	disorder
2025	like	anxiety	thought	disorder	think

Table 21: Top-5 Co-occurring Neighbors per Year for *harassment* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	sexual	get	send	sexually	research
2011	student	sexual	use	like	people
2012	like	rape	tell	sexual	people
2013	think	child	people	man	woman
2014	sexual	people	bullying	likely	tell
2015	woman	female	think	online	sexual
2016	people	like	bullying	health	mental
2017	people	sexual	woman	man	like
2018	sexual	woman	people	like	workplace
2019	people	like	know	want	talk
2020	people	like	sexual	sexually	sub
2021	sexual	people	like	feel	think
2022	people	sexual	woman	man	sexually
2023	act	insult	tolerate	provoke	language
2024	act	provoke	insult	disrespectfully	racist
2025	insult	act	racist	language	tolerate

Table 25: Top-5 Co-occurring Neighbors per Year for *prejudice* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	see	know	self	yes	fight
2011	people	aware	mean	study	find
2012	people	think	social	mental	way
2013	people	think	racism	mean	say
2014	implicit	people	attitude	association	base
2015	people	find	black	base	study
2016	people	research	like	think	bias
2017	people	think	racism	feel	study
2018	people	group	study	right	like
2019	people	group	study	right	abstract
2020	people	like	mental	group	woman
2021	people	group	like	bias	think
2022	people	racism	like	group	bias
2023	people	woman	like	man	experience
2024	people	think	like	bias	man
2025	people	group	like	man	social

Table 22: Top-5 Co-occurring Neighbors per Year for *imposter* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	delusion	syndrome(http://en.wikipedia.org)	wiki	impostor_syndrome	interesting
2011	right	people	problem	convince	close
2012	delusion	person	capgras	replace	see
2013	syndrome	research	effect	real	actually
2014	syndrome	control	simply	forget	picture
2015	syndrome	develop	think	people	feel
2016	syndrome	think	like	people	call
2017	syndrome	delusion	like	opposite	replace
2018	syndrome	school	grad	life	feel
2019	syndrome	like	think	feel	capable
2020	like	syndrome	feel	think	time
2021	syndrome	like	feel	people	know
2022	syndrome	like	feel	think	people
2023	syndrome	feel	like	feeling	people
2024	like	syndrome	feel	people	try
2025	syndrome	like	feel	people	know

Table 26: Top-5 Co-occurring Neighbors per Year for *psychopath* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	people	think	like	disorder	personality
2011	people	think	like	empathy	trait
2012	sociopath	people	disorder	think	like
2013	people	think	disorder	sociopath	personality
2014	people	like	think	sociopath	disorder
2015	people	think	disorder	like	sociopath
2016	people	think	personality	sociopath	like
2017	people	like	disorder	think	personality
2018	people	trait	personality	like	disorder
2019	people	trait	think	personality	disorder
2020	people	sociopath	like	think	disorder
2021	people	like	narcissistic	think	personality
2022	people	like	think	disorder	empathy
2023	people	narcissistic	like	personality	trait
2024	people	like	narcissistic	think	trait
2025	people	trait	narcissistic	like	empathy

Table 27: Top-5 Co-occurring Neighbors per Year for *toxic* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	people	deficiency	and/or	like	environment
2011	environment	deficiency	brain	good	different
2012	people	individual	help	deficiency	brain
2013	change	social	health	mean	behavior
2014	drug	like	people	brain	cause
2015	people	like	article	feel	think
2016	masculinity	man	people	like	think
2017	masculinity	people	like	man	life
2018	people	relationship	like	life	mental
2019	people	like	relationship	masculinity	think
2020	people	like	relationship	feel	life
2021	people	like	relationship	think	feel
2022	people	like	relationship	thing	think
2023	people	like	man	relationship	masculinity
2024	people	like	relationship	man	masculinity
2025	people	like	man	relationship	woman

Table 28: Top-5 Co-occurring Neighbors per Year for *trauma* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	event	child	experience	childhood	abuse
2011	experience	event	like	stress	people
2012	experience	memory	event	people	like
2013	experience	people	like	cause	child
2014	experience	people	childhood	like	ptsd
2015	experience	people	think	event	ptsd
2016	experience	people	ptsd	like	disorder
2017	psychotherapy	processing	like	sensorimotor	sp4
2018	experience	people	like	childhood	help
2019	experience	people	like	childhood	help
2020	experience	like	people	feel	childhood
2021	like	experience	people	help	feel
2022	experience	like	people	childhood	therapist
2023	experience	people	like	therapist	feel
2024	sexual	experience	help	like	recover
2025	sexual	violence	rainn.org](https://rainn.org	recover	help

Table 29: Top-5 Co-occurring Neighbors per Year for *triggered* - Reddit Psychology Corpus

Year	1	2	3	4	5
2010	think	brain	like	fear	way
2011	like	people	time	brain	thing
2012	like	people	thing	think	stress
2013	thing	people	like	anxiety	know
2014	like	thing	people	help	know
2015	people	warming	like	think	help
2016	like	anxiety	thing	feel	people
2017	like	people	thing	think	feel
2018	like	thing	people	anxiety	feel
2019	like	feel	thing	people	anxiety
2020	like	thing	feel	people	know
2021	like	feel	thing	people	anxiety
2022	like	people	feel	thing	know
2023	like	feel	thing	trauma	people
2024	get	give	topic	provide	potentially
2025	prevent	resource	get	give	topic

Table 30: Top-5 Co-occurring Neighbors per Year for *addiction* - Reddit General Corpus

Year	1	2	3	4	5
2010	drug	like	people	time	get
2011	drug	people	think	like	know
2012	drug	people	like	money	welfare
2013	drug	like	people	think	get
2014	drug	people	like	think	get
2015	drug	like	think	people	life
2016	drug	like	people	get	think
2017	drug	like	year	people	life
2018	drug	people	like	life	get
2019	drug	like	sex	people	thing
2020	drug	porn	like	people	year
2021	drug	like	porn	people	year
2022	drug	porn	like	life	people
2023	drug	people	like	life	porn
2024	drug	porn	like	people	time
2025	drug	porn	like	people	time

Table 31: Top-5 Co-occurring Neighbors per Year for *bipolar* - Reddit General Corpus

Year	1	2	3	4	5
2010	disorder	people	know	like	medication
2011	disorder	manic	people	think	depression
2012	disorder	like	diagnose	know	depressed
2013	disorder	people	like	depression	mom
2014	disorder	people	depression	anxiety	diagnose
2015	disorder	depression	time	year	like
2016	disorder	people	diagnose	depression	get
2017	disorder	depression	think	thing	work
2018	disorder	year	like	depression	diagnose
2019	disorder	depression	diagnose	mental	people
2020	disorder	depression	people	diagnose	date
2021	disorder	diagnose	like	depression	year
2022	disorder	people	diagnose	depression	mental
2023	disorder	diagnose	like	year	get
2024	disorder	feel	like	med	diagnose
2025	disorder	depression	get	like	diagnose

Table 32: Top-5 Co-occurring Neighbors per Year for *bullying* - Reddit General Corpus

Year	1	2	3	4	5
2010	kid	school	get	like	people
2011	school	kid	get	people	like
2012	school	kid	get	like	people
2013	school	kid	people	get	like
2014	school	people	friend	kid	like
2015	school	people	like	kid	get
2016	school	kid	get	people	like
2017	school	people	kid	like	think
2018	school	kid	like	people	get
2019	school	like	people	get	friend
2020	school	people	get	like	friend
2021	school	people	get	like	year
2022	school	like	people	get	kid
2023	school	people	get	like	kid
2024	school	get	like	people	friend
2025	school	kid	get	people	like

Table 33: Top-5 Co-occurring Neighbors per Year for *gaslighting* - Reddit General Corpus

Year	1	2	3	4	5
2010	band	anthem	anthem](http://www.youtube.com	watch?v	current
2011	wiki	family	pretty	bit	let
2012	feel	anthem	begin	thing	time
2013	ray	lamontagne	jolene	mike	doughty
2014	gaslighte	man	behavior	make	victim
2015	later	know	people	woman	break
2016	year	emotional	horrible	relationship	time
2017	like	try	abuse	know	time
2018	try	lie	abuse	think	emotional
2019	like	woman	tell	crazy	bullying
2020	try	people	abuse	thing	want
2021	like	try	think	abuse	tell
2022	try	person	people	manipulate	relationship
2023	try	people	think	like	tell
2024	woman	try	like	people	thing
2025	try	think	lie	like	woman

Table 34: Top-5 Co-occurring Neighbors per Year for *harassment* - Reddit General Corpus

Year	1	2	3	4	5
2010	people	woman	sexual	get	like
2011	sexual	people	get	like	ask
2012	people	sexual	like	guy	tell
2013	sexual	woman	man	people	like
2014	sexual	woman	like	man	people
2015	woman	sexual	like	people	man
2016	woman	sexual	people	man	like
2017	sexual	woman	like	guy	get
2018	sexual	woman	man	people	assault
2019	woman	sexual	man	like	think
2020	sexual	woman	man	like	sexually
2021	sexual	woman	man	sexually	like
2022	woman	sexual	sexually	man	people
2023	woman	man	sexual	people	get
2024	woman	sexual	man	sexually	like
2025	woman	sexual	man	get	people

Table 38: Top-5 Co-occurring Neighbors per Year for *prejudice* - Reddit General Corpus

Year	1	2	3	4	5
2010	people	think	like	way	come
2011	people	know	minority	say	stereotype
2012	people	think	white	racist	like
2013	people	think	woman	man	pride
2014	people	like	think	man	know
2015	people	like	think	man	woman
2016	pride	people	like	thing	woman
2017	people	racism	white	pride	like
2018	people	think	like	woman	experience
2019	people	woman	like	think	pride
2020	stereotype	personal	experience	opinion	entire
2021	experience	preference	opinion	personal	stereotype
2022	personal	experience	generalize	opinion	stereotype
2023	people	experience	personal	stereotype	opinion
2024	preference	generalize	opinion	experience	question
2025	people	like	man	pride	woman

Table 35: Top-5 Co-occurring Neighbors per Year for *imposter* - Reddit General Corpus

Year	1	2	3	4	5
2010	lose	doubt	wait	lie	hat
2011	syndrome	artist	know	think	lose
2012	dream	lose	meet	man	triumph
2013	like	lose	man	doubt	wait
2014	lose	man	life	think	thing
2015	syndrome	like	feel	look	thing
2016	syndrome	man	lose	hold	doubt
2017	man	lose	turn	doubt	wait
2018	man	syndrome	feel	like	know
2019	like	syndrome	feel	work	good
2020	syndrome	like	feel	job	woman
2021	syndrome	like	feel	think	smart
2022	syndrome	feel	like	need	bisexual
2023	syndrome	feel	like	man	dream
2024	syndrome	feel	like	work	time
2025	syndrome	like	wait	know	lose

Table 39: Top-5 Co-occurring Neighbors per Year for *psychopath* - Reddit General Corpus

Year	1	2	3	4	5
2010	think	like	people	want	come
2011	people	like	think	personality	believe
2012	like	disorder	people	personality	indicate
2013	people	like	think	know	girl
2014	like	people	woman	guy	way
2015	like	people	think	emotional	want
2016	people	think	get	like	know
2017	think	like	people	person	know
2018	like	people	say	thing	feel
2019	like	people	check	know	bad
2020	people	like	job	think	sound
2021	like	people	think	know	thing
2022	people	like	think	narcissistic	want
2023	like	people	man	sociopath	think
2024	like	people	way	play	try
2025	people	think	narcissistic	like	woman

Table 36: Top-5 Co-occurring Neighbors per Year for *narcissistic* - Reddit General Corpus

Year	1	2	3	4	5
2010	think	people	self	like	good
2011	think	like	personality	people	say
2012	like	people	look	think	say
2013	like	people	feel	want	thing
2014	think	people	like	self	woman
2015	like	people	self	think	woman
2016	like	think	people	personality	disorder
2017	people	think	like	good	want
2018	like	people	think	personality	mother
2019	like	people	thing	think	life
2020	people	like	self	think	get
2021	people	like	think	year	mother
2022	people	think	like	good	know
2023	year	people	abusive	abuse	like
2024	people	woman	want	like	think
2025	people	like	think	life	woman

Table 40: Top-5 Co-occurring Neighbors per Year for *toxic* - Reddit General Corpus

Year	1	2	3	4	5
2010	chemical	like	find	say	people
2011	like	body	waste	people	relationship
2012	know	substance	relationship	like	thing
2013	relationship	like	know	think	people
2014	relationship	people	like	think	thing
2015	woman	people	man	think	relationship
2016	people	relationship	friend	life	masculinity
2017	relationship	people	like	life	get
2018	relationship	people	masculinity	like	man
2019	relationship	people	masculinity	like	friend
2020	relationship	people	masculinity	like	friend
2021	people	relationship	like	think	friend
2022	health	mental	people	term	label
2023	health	mental	people	man	term
2024	health	mental	people	relationship	woman
2025	people	health	mental	relationship	man

Table 37: Top-5 Co-occurring Neighbors per Year for *ocd* - Reddit General Corpus

Year	1	2	3	4	5
2010	like	people	thing	think	disorder
2011	like	people	thing	get	life
2012	people	get	thing	like	know
2013	like	people	thing	think	know
2014	anxiety	like	thing	issue	people
2015	like	thing	anxiety	think	need
2016	anxiety	depression	like	time	mental
2017	anxiety	like	depression	disorder	thing
2018	like	anxiety	people	thing	depression
2019	anxiety	like	thing	depression	bad
2020	anxiety	like	thing	think	lot
2021	anxiety	like	get	thing	adhd
2022	like	thing	anxiety	think	time
2023	like	thing	issue	help	anxiety
2024	get	like	thought	thing	people
2025	like	disorder	people	anxiety	adhd

Table 41: Top-5 Co-occurring Neighbors per Year for *trauma* - Reddit General Corpus

Year	1	2	3	4	5
2010	experience	time	way	thing	think
2011	experience	people	child	like	life
2012	people	cause	experience	life	get
2013	like	experience	people	think	cause
2014	experience	think	know	people	like
2015	people	experience	like	think	thing
2016	experience	like	life	people	think
2017	experience	think	people	like	life
2018	experience	like	people	think	life
2019	like	experience	feel	people	think
2020	permit	tourism	askwomen	comments8xqo	announcementupdate
2021	people	like	experience	lot	childhood
2022	like	think	childhood	year	experience
2023	people	like	life	time	experience
2024	like	people	childhood	thing	past
2025	like	people	man	lot	think

Table 42: Top-5 Co-occurring Neighbors per Year for *triggered* - Reddit General Corpus

Year	1	2	3	4	5
2010	pull	time	like	thing	know
2011	pull	like	gun	people	time
2012	pull	know	get	gun	people
2013	pull	like	time	thing	think
2014	like	pull	know	think	thing
2015	pull	like	time	people	thing
2016	like	think	pull	thing	time
2017	post	error	chance	violate	rule
2018	like	anxiety	know	get	people
2019	chrono	like	time	know	thing
2020	like	thing	people	pull	get
2021	like	time	get	thing	pull
2022	like	feel	get	thing	know
2023	like	thing	people	know	get
2024	like	thing	time	know	pull
2025	like	thing	people	get	know

Rethinking Metrics for Lexical Semantic Change Detection

Roksana Goworek

Queen Mary University of London
r.goworek@qmul.ac.uk

Haim Dubossarsky

Queen Mary University of London
The Alan Turing Institute
University of Cambridge
h.dubossarsky@qmul.ac.uk

Abstract

Lexical semantic change detection (LSCD) increasingly relies on contextualised language model embeddings, yet most approaches still quantify change using a small set of semantic change metrics, primarily Average Pairwise Distance (APD) and cosine distance over word prototypes (PRT). We introduce Average Minimum Distance (AMD) and Symmetric Average Minimum Distance (SAMD), new measures that quantify semantic change via local correspondence between word usages across time periods. Across multiple languages, encoder models, and representation spaces, we show that AMD often provides more robust performance, particularly under dimensionality reduction and with non-specialised encoders, while SAMD excels with specialised encoders. We suggest that LSCD may benefit from considering alternative semantic change metrics beyond APD and PRT, with AMD offering a robust option for contextualised embedding-based analysis.

1 Introduction

Lexical Semantic Change Detection (LSCD) aims to automatically quantify how word meanings evolve over time through the analysis of diachronic text corpora (Tahmasebi and Dubossarsky, 2023). Recent shared tasks and benchmarks have established a standard evaluation setting in which systems must determine, for a given target word and two time periods, whether its meaning has changed and to what extent (Schlechtweg et al., 2020a, 2024). With the rise of contextualised language models, LSCD approaches have increasingly moved from static type-level representations to usage-level embeddings, in which each occurrence of a target word is encoded

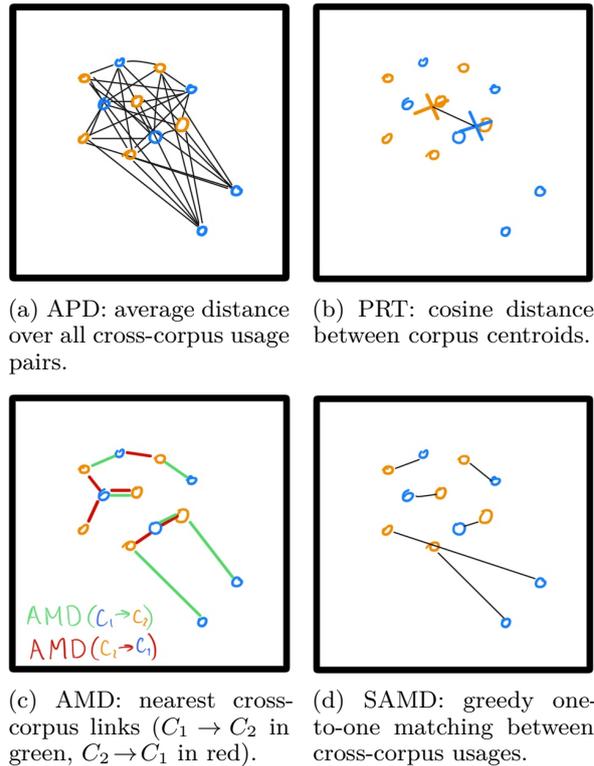


Figure 1: Illustrative schematic contrasting LSCD distance measures on usage embeddings from two corpora (blue: C_1 , yellow: C_2).

as a high-dimensional vector reflecting its local context (Martinc et al., 2020). Semantic change is then quantified by comparing the distributions of these usage embeddings across corpora, most commonly using semantic change metrics such as Average Pairwise Distance (APD) or distances between corpus-specific word prototypes (PRT) (Kutuzov and Giulianelli, 2020a).

While these approaches achieve strong performance on benchmark datasets, the choice of semantic change metric implicitly encodes assumptions about what constitutes semantic change. Metrics such as APD and PRT

aggregate information globally—either across all cross-corpus usage pairs or via a single centroid per corpus—which makes them well suited to capturing broad distributional shifts. This global aggregation can obscure localised phenomena, such as the emergence of a new sense, the disappearance of an older one, or changes affecting only a subset of usages whose occurrences might be too rare to notice.

In this paper, we introduce **Average Minimum Distance (AMD)** and its **Symmetric** variant **SAMD**, a pair of semantic change metrics based on usage-level nearest cross-corpus correspondence. For each usage in one corpus, AMD measures the distance to its closest counterpart in the other corpus and averages these minimum distances in both directions, while SAMD further enforces a one-to-one alignment between usage instances across periods. This formulation is grounded in the intuition that semantic change is reflected in how well individual usages from one corpus can be matched to semantically similar usages in another, rather than solely in how far two usage distributions are on average. As illustrated in [Figure 1](#), AMD and SAMD emphasise local correspondence between usages and naturally capture usage-level changes.

AMD permits a directional decomposition, distinguishing how well earlier-corpus usages are explained by later-corpus usages and vice versa. This makes it possible to analyse asymmetric changes such as broadening/narrowing or sense emergence and disappearance. Moreover, because AMD and SAMD rely on usage-level correspondences rather than global distributional geometry, they can be applied more meaningfully in reduced or restructured representation spaces.

Our experiments show that AMD and SAMD provide robust and complementary alternatives to existing semantic change metrics, particularly when embeddings are projected into lower-dimensional spaces, or when non-specialised encoders are used. We release code, results, and generated definitions to reproduce our experiments and construct definition-based spaces.¹

¹https://github.com/roksanagow/Rethinking_LSCD_Metrics

2 Related Work

2.1 Semantic change metrics for LSCD

Early work on LSCD was developed in the context of static word embeddings, where a word is represented by a single vector per corpus (Kim et al., 2014; Hamilton et al., 2016a; Schlechtweg et al., 2019). In this setting, semantic change is naturally quantified by comparing word representations across corpora. With the introduction of contextualised models this paradigm was reintroduced using PRT, which averages the representations of all usages of a particular word, leaving a single vector for subsequent analysis as in static models.

In contrast, global-distribution-based metrics such as Average Pairwise Distance (APD) are motivated by the availability of contextualised usage embeddings and explicitly compare distributions of individual word occurrences across corpora (Kutuzov and Giulianelli, 2020b; Kutuzov et al., 2022b). By exploiting the richer structure of contextualised representations, APD extends prototype-based approaches and has become a standard usage-level semantic change metric in contemporary LSCD benchmarks, usually at the top of the leader-board (Periti and Tahmasebi, 2024).

2.2 Robustness and discovery-oriented LSCD

Recent work has raised concerns about whether strong benchmark performance translates into reliable semantic-change discovery in realistic settings (Hamilton et al., 2016b; Kurtyigit et al., 2021). In particular, Umarova et al. (2025) show that when LSCD systems are applied to full vocabularies, many high-scoring words correspond to topical or domain variation rather than genuine semantic change. This finding highlights a gap between curated benchmark evaluations and large-scale discovery scenarios, and suggests that semantic change metrics emphasising global distributional shifts may be vulnerable to confounds when applied broadly (Schlechtweg et al., 2019; Dubossarsky et al., 2019).

2.3 Interpretable LSCD

In parallel, several strands of research have explored more interpretable frameworks for

LSCD. Hypothesis-driven approaches incorporate expert-defined sense distinctions to track fine-grained semantic change, arguing that unsupervised methods may fail to align representations with linguistically meaningful senses (Cassotti et al., 2023). Fedorova et al. (2024) propose generating LLM-based definitions for individual usages and detecting change by comparing distributions of induced sense labels, while SCDTour (Aida and Bollegala, 2025) identifies interpretable axes of change in static embedding spaces. These approaches demonstrate the potential of interpretable representations, but rely on specialised pipelines or additional modelling assumptions, and typically retain global semantic change metrics such as APD or PRT for quantification.

3 Method

3.1 Task Formulation and Notation

We follow the standard lexical semantic change detection (LSCD) setting. For a given language, we are given two diachronic corpora C_1 and C_2 , corresponding to an earlier and a later time period. Let \mathcal{W} denote the set of target lemmas for this language. For each $w \in \mathcal{W}$ and each corpus C_t ($t \in \{1, 2\}$), we extract all sentences (or sentence-like contexts) that contain an occurrence of w . We write

$$U_t(w) = \{u_1^{(t)}, \dots, u_{n_t}^{(t)}\}$$

for the set of usage instances of w in period t , where each $u_i^{(t)}$ is a sequence of tokens with a marked target position.

Let f_θ be a contextual encoder that maps a usage u with a marked target token to a contextual embedding $v = f_\theta(u) \in \mathbb{R}^D$. We denote the resulting sets of usage embeddings as

$$V_t(w) = \{v_1^{(t)}, \dots, v_{n_t}^{(t)}\} \subset \mathbb{R}^D.$$

For notational simplicity, in the remainder of this section we write $A = V_1(w)$ and $B = V_2(w)$ for the sets of usage embeddings of a target word w in the two time periods. Given A and B , we compute a graded semantic change score using the LSCD metrics described in subsection 3.2.

Evaluation. For evaluation, each target word is assigned a single semantic change score by a given metric, and performance is measured as

the Spearman rank correlation between these scores and the gold graded change scores provided with the datasets, averaged across all target words.

Our method is encoder-agnostic: it operates on any token-level embedding model f_θ , and we compare it across multiple encoders described in subsection 3.3.

Datasets. We evaluate on LSCD benchmarks in seven languages: English, German, Swedish, and Latin datasets from SemEval-2020 Task 1 (Schlechtweg et al., 2020b), the Spanish semantic change dataset from LSCDiscovery (Zamora-Reina et al., 2022), the Norwegian dataset NorDiaChange (Kutuzov et al., 2022a), and the Chinese semantic shift benchmark (Chen et al., 2022) for Chinese. Each benchmark provides diachronic corpus pairs with human-annotated semantic change scores.

3.2 LSCD Metrics

All metrics use cosine distance, defined as $\delta(x, y) = 1 - \cos(x, y)$.

Average Pairwise Distance (APD) computes the mean distance between all cross-corpus usage pairs:

$$\text{APD}(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} \delta(a, b).$$

APD captures global divergence between usage distributions but is dominated by the bulk of points, making it less sensitive to small or emerging usage clusters.

Prototype Distance (PRT) compares the centroids of the two periods. Let

$$\mu(A) = \frac{1}{|A|} \sum_{a \in A} a.$$

Then

$$\text{PRT}(A, B) = \delta(\mu(A), \mu(B)).$$

PRT is computationally efficient but reduces the representations of a word in each period to a single point, potentially obscuring sense-specific changes.

Average Minimum Distance (AMD). We introduce AMD, a nearest-neighbour-based metric designed to capture local discrepancies between periods. For a point x and a set Y , we define

$$\text{nn-dist}(x, Y) = \min_{y \in Y} \delta(x, y).$$

The directional score is

$$\text{AMD}(A \rightarrow B) = \frac{1}{|A|} \sum_{a \in A} \text{nndist}(a, B),$$

with $\text{AMD}(B \rightarrow A)$ defined analogously. The symmetric score is

$$\text{AMD}(A, B) = \frac{\text{AMD}(A \rightarrow B) + \text{AMD}(B \rightarrow A)}{2}.$$

Symmetric Average Minimum Distance (SAMD). While AMD allows many-to-one nearest-neighbour mappings, we further introduce SAMD, which enforces one-to-one correspondence between usage instances across periods. We sample an equal number of usages from both corpora, corresponding to the minimum available across the two periods, such that $|A| = |B|$. Define the pairwise distances $D_{ij} = \delta(a_i, b_j)$. SAMD greedily selects the smallest remaining distance, removes the corresponding row and column, and repeats until all usage pairs are matched.² The final score is

$$\text{SAMD}(A, B) = \frac{1}{|A|} \sum_{k=1}^{|A|} \delta(a_{i_k}, b_{j_k}).$$

AMD captures local shifts by matching each usage to its closest counterpart, but may map multiple points onto the same neighbour. This behaviour can be problematic in highly hub-dominated spaces (Radovanović et al., 2010; Dinu and Baroni, 2014) (see Appendix D for an analysis of hubness in the selected embedding spaces). This phenomenon is of particular concern in LSCD, where a small number of modern-style usages in the historical corpus may emerge as nearest neighbours for a disproportionate number of modern usages, despite lacking semantic correspondence. SAMD mitigates this effect by approximating a minimal one-to-one alignment between periods, yielding a metric that is sensitive to distributional shifts as well as cross-corpus correspondence.

3.3 Models

We evaluate AMD across a diverse set of pre-trained language models in order to assess its robustness. We include **XL-LEXEME**

²We also evaluated optimal one-to-one matching via the Hungarian algorithm and found results to be near-identical to the greedy algorithm used here.

(Cassotti et al., 2023), a multilingual encoder, based on XLM-RoBERTa, explicitly fine-tuned for word-in-context (WiC) discrimination. XL-LEXEME has been shown to achieve state-of-the-art performance on LSCD.

Multilingual encoders. We compare several widely used multilingual encoders without any fine-tuning, including XLM-RoBERTa (Conneau et al., 2020), a strong general-purpose multilingual baseline; the recently released mmBERT (Marone et al., 2025), designed for improved multilingual semantic alignment; RemBERT (Chung et al., 2021), which emphasises language-balanced pretraining across many languages; and multilingual-E5 (Wang et al., 2024), a retrieval-oriented model optimised for representing semantic similarity. These models are commonly employed in cross-lingual and multilingual semantic tasks, and allow us to test whether different LSCD metrics remain effective when applied to generic contextual representations.

Monolingual encoders. We also evaluate various monolingual encoders, selecting one model per language: RoBERTa for English (Liu et al., 2019), GBERT for German (Chan et al., 2020), BammanBurns-BERT for Latin (Bamman and Burns, 2020), Spanish BERT (Cañete et al., 2023), Chinese RoBERTa (Cui et al., 2020), NB-BERT for Norwegian (Kummervold et al., 2021), and Megatron-BERT for Swedish (Malmsten et al., 2020). These models are trained on large language-specific corpora and provide strong language-specific contextual representations for LSCD. Focusing on monolingual encoders is meant to test AMD and SAMD under conditions mimicking low resource (and historical) languages (e.g., Latin, Ancient Greek), or on specialized domains (e.g., medical, legal), where multilingual models struggle or may not be available.

3.4 Reduced Space Construction

To analyse the behaviour of LSCD metrics under varying representational constraints, we evaluate them not only on the original contextual embedding space, but also on several reduced-dimensional spaces constructed from the same usage embeddings. These spaces preserve different aspects of the original representations: (i) full embeddings, which retain

all information; (ii) a definition-based space, which preserves semantically interpretable information; (iii) principal components, which preserve maximal global variance; and (iv) randomly selected dimensions, which preserve neither semantic structure nor global variance in any principled way.

All LSCD metrics are applied either to the original sets $A, B \subset \mathbb{R}^D$ or to their transformed counterparts $\phi(A), \phi(B) \subset \mathbb{R}^K$, depending on the representation space.

Definition-space. For each target word w , we generate a finite set of textual definitions

$$\mathcal{D}(w) = \{d_1, \dots, d_K\}$$

using **Gemini 2.5 Pro** (Comanici et al., 2025), a multilingual large language model. The same prompt, translated into each target language, is used across languages (see Table 2 for prompts and Table 1 for the amounts of generated definitions). We do not sample definitions from existing dictionaries, as it is difficult to obtain resources that provide multiple definitions consistently for all target words. Moreover, dictionary entries for different words are often authored by different lexicographers, which can introduce variation in sense granularity and segmentation across entries. Our setup instead mirrors a scenario in which a single lexicographer provides a set of candidate definitions per investigated word.

Each definition is prepended with the target word, yielding inputs of the form $w: d_k$ where d_k is the k^{th} definition for word w . We obtain contextualised embeddings for the target word within each definition using the same encoder f_θ used to encode the usages, resulting in definition embeddings

$$z_k = f_\theta(w: d_k) \in \mathbb{R}^D.$$

Given a usage embedding $v \in A \cup B$, we project it into definition-space by computing its cosine distance to each definition embedding:

$$\phi_{\text{def}}(v) = (\delta(v, z_1), \dots, \delta(v, z_K)) \in \mathbb{R}^K.$$

The resulting vectors are treated as standard embeddings in \mathbb{R}^K , to which APD, PRT, AMD and SAMD are applied directly.

Principal component analysis (PCA).

As a variance-preserving but non-interpretable projections, we apply PCA independently for each target word, using all its usage embeddings across both periods. Each embedding $v \in \mathbb{R}^D$ is projected onto the top K principal components. Unless stated otherwise, K matches the number of definitions used in the definition-space projection for direct comparison, while the stress test varies K to assess sensitivity to number of dimensions.

Random dimension selection. An extreme case of dimensionality is introduced by randomly selecting K embedding dimensions independently for each target word. The same dimensions are retained for all usages of that word across both periods, allowing us to isolate the effect of dimensionality reduction from that of principled subspace construction. Like for PCA, K matches the number of definitions generated for a given word to facilitate comparison, while the stress test varies K to assess sensitivity of metrics to number of dimensions.

4 Results

We evaluate LSCD metrics across representation spaces, encoder families, and languages, with a focus on their robustness under varying representational constraints. All results report Spearman correlations with graded semantic change scores and are averaged across languages and encoder models unless stated otherwise.



Figure 2: Performance (Spearman correlations) across metrics and spaces, averaged over languages and encoders. Std in brackets.

Overall comparison. Figure 2 presents a high-level comparison of APD, PRT, AMD, and SAMD across full embeddings and reduced-dimensional spaces. The results show a consistent pattern: correspondence-based metrics have the best performance (in terms of average Spearman correlations) and the most stable (in terms of lowest variance) across all four types of embedding spaces (FULL, DEF, PCA, RAND).

The best performance of AMD and SAMD is when they are paired with PCA dimensionality reduction, suggesting that PCA may reduce irrelevant noise in the embeddings. Notably, AMD and SAMD can uniquely withstand the noise elicited by the more aggressive dimensionality reduction methods (PCA and RAND), which can even improve performance.

Moreover, AMD is the only metric that significantly gains from projecting its representations to the definition space, while APD and PRT show a significant drop in performance. In the interpretable definition-based space, AMD achieves the highest average performance, suggesting that, in this space, semantic change is better captured through differences in the alignment of the overall shape of usage embedding distributions, rather than through shifts in their distribution concentration. In contrast, AMD and SAMD perform similarly in FULL, PCA and RAND spaces, while remaining markedly more robust than APD and PRT.

Overall, these results demonstrate that local correspondence metrics provide both stronger and more stable estimates of semantic change across diverse representation spaces.



Figure 3: Performance of APD, PRT, AMD and SAMD using XL-LEXEME embeddings across representation spaces. Each row corresponds to a language; columns show metric–space combinations.

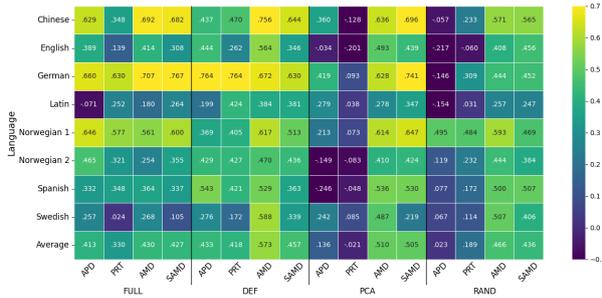
Specialised encoder: XL-LEXEME. We next examine results obtained with XL-LEXEME, a model explicitly fine-tuned for LSCD (Figure 3). Here too the best performing approach is SAMD with PCA, achieving an average Spearman correlation of .694 across all languages, surpassing APD (.635) and PRT (.643) on full embeddings. SAMD also outperforms standard metrics in the full embedding space.

In the definition-based space, AMD yields the highest correlation (.640), again highlighting the effectiveness of directional nearest-neighbour alignment in structured semantic spaces. However, the gains from combining AMD with DEF space are markedly larger for non-specialised encoders than for XL-LEXEME, likely because XL-LEXEME’s full embedding space is already organised in a sense-aware fashion through its fine-tuning.

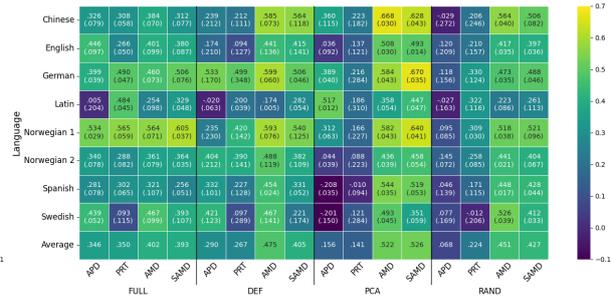
As observed in the aggregate analysis in Figure 2, both AMD and SAMD remain robust under PCA and random dimensionality reduction, whereas APD and PRT exhibit detrimental degradation. These findings indicate that even for specialised encoders, local correspondence metrics provide robust estimates of semantic change, especially under dimensionality reduction.

Non-specialised encoders. Figures 4a and 4b summarise results for language-specific monolingual encoders and general-purpose multilingual encoders, respectively. Across both settings, performance is lower than for specialised encoders. Nevertheless, the benefit of our two novel metrics is clear. Notably, monolingual encoders consistently outperform the averaged multilingual models, suggesting that language-specific models may be better suited to capturing fine-grained usage distinctions. This aligns with prior findings questioning the effectiveness of multilingual representations, or multilingual fine-tuning for sense-level tasks (Goworek and Dubossarsky, 2025; Goworek et al., 2025). Fine-tuning such monolingual encoders on WiC may therefore yield stronger performance than relying on a single general-purpose multilingual model.

For monolingual encoders (Figure 4a), AMD paired with the definition-based space yields the strongest average performance, substan-



(a) Monolingual encoders (one per language).



(b) Multilingual encoders (averaged over models; std).

Figure 4: Performance of APD, PRT, AMD and SAMD for non-specialised encoders.

tially outperforming all other metric-space combinations. This could suggest that the sense-aware organisation of the embedding space is conducive to highlighting change which manifests as difficult-to-pair usages across corpora. The next best results are achieved by AMD and SAMD under PCA reduction, supporting previous results. On full embeddings, the correspondence-based metrics again outperform APD and PRT, and these global-shift metrics collapse under PCA and random reduction, often approaching zero correlation.

A similar pattern emerges for multilingual encoders (Figure 4b). The best average performance is obtained by SAMD and AMD under PCA reduction (.526 and .522), followed by AMD in the definition space (.475). One possible explanation for the weaker benefit of the AMD-DEF combination compared to monolingual models is that multilingual encoders may exhibit weaker sense-level organisation within each language, as they must represent a much broader multilingual vocabulary, making semantic projections less effective.

In the full space, AMD and SAMD remain around .400, while APD and PRT achieve a Spearman correlation of approximately .350. Again, APD and PRT exhibit severe sensitivity to dimensionality reduction.

Taken together, these results confirm that AMD and SAMD provide substantially more robust semantic change estimates than global metrics, particularly in constrained or lower-dimensional representation spaces.

4.1 Stress Test

To further analyse robustness, we progressively reduce embedding dimensionality by fac-

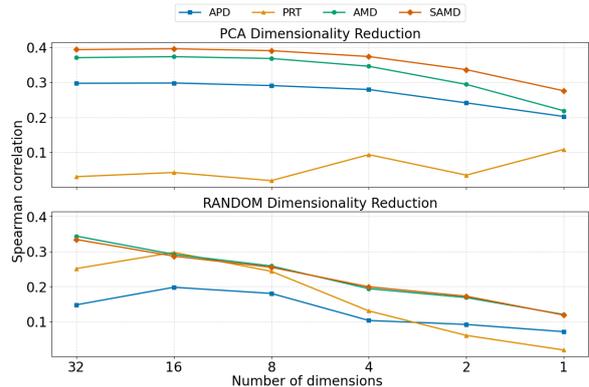


Figure 5: Stress test of metrics under progressive dimensionality reduction. Spearman correlations are averaged across languages and encoders as embedding dimensionality is reduced by factors of two using PCA or random dimension selection.

tors of two using PCA and random dimension selection (Figure 5). PRT fails almost completely under PCA, yielding correlations close to zero even at moderate dimensionalities. APD yields average correlations around .300 in low-dimensional spaces. In contrast, AMD and SAMD exhibit greater robustness, maintaining correlations near .400 up until only 4 dimensions are retained. Under PCA, SAMD shows slightly higher robustness than AMD, while under random reduction both metrics perform almost identically. Across most numbers of dimensions, APD and PRT remain consistently worse by almost .100. These results highlight that global distribution-based metrics rely heavily on high-dimensional embedding structure, whereas local correspondence metrics retain meaningful performance under dimensionality reduction.

5 Discussion

Our results show that while some metrics are, on average, better than others, their effectiveness largely depends on encoder choices (e.g., multi or monolingual, finetuned or not) as well as on the embedding quality and representational constraints (e.g., dimensionality reduction or semantic projection). This aligns with recent work showing that LSCD performance is highly sensitive to modelling choices and evaluation conditions, with no single method consistently dominating across settings or languages (Periti and Tahmasebi, 2024).

This is also supported by work showing that LSC itself may manifest along different dimensions of change such as affective or pragmatic shifts, and that no single model is found to be the best across different change types (Baes et al., 2025; Goworek and Dubossarsky, 2024).

Local correspondence vs. global aggregation. Prototype-based measures such as PRT and global usage-level metrics such as APD summarise semantic change by aggregating information across all usages. As a result, they are most sensitive to changes that affect the overall distribution of usages or their centroid, making them effective when semantic change manifests as a broad redistribution of usage embeddings. This behaviour has been noted in prior work, where APD and PRT were shown to exhibit dataset-dependent performance, with preferences for one over the other correlating with properties of the target-word change score distributions rather than a single metric consistently dominating across conditions (Kutuzov et al., 2022b; Martinc et al., 2020).

In contrast, the focus of AMD and SAMD on local correspondence between usages makes them particularly sensitive to changes affecting only a subset of usages, which may be diluted by global aggregation. While AMD allows multiple usages to map to the same nearest neighbour and can be decomposed into more interpretable directional components, SAMD enforces a one-to-one alignment between periods, making it sensitive not only to changes in distributional shape but also to shifts in the concentration of usage embeddings. These observations suggest that semantic change metrics are not interchangeable.

Their behaviour differs systematically depending on how they aggregate information across usages and how they interact with representational constraints (Kutuzov et al., 2022b).

Robustness and dimensions. Our results highlight that semantic change metrics differ substantially in their sensitivity to representational quality and embedding structure. Global distribution-based metrics such as APD and PRT degrade sharply under dimensionality reduction and are less optimal with less specialised embedding spaces, whereas local correspondence-based metrics such as AMD and SAMD remain considerably more robust.

This finding resonates with earlier observations that LSCD performance is strongly influenced by embedding alignment quality and representational noise (Martinc et al., 2020; Giulianelli et al., 2020). It also complements recent work questioning the benefits of multilingual fine-tuning and representations for fine-grained semantic tasks, where monolingual model often outperform general-purpose multilingual models, even after fine-tuning (Goworek and Dubossarsky, 2025; Goworek et al., 2025).

Our findings may additionally connect to concerns about large-scale semantic-change discovery (Umarova et al., 2025). Methods optimised for curated benchmark evaluations frequently prioritise detecting broad semantic shift over more realistic and nuanced semantic change when applied across full vocabularies. By focusing on usage-level correspondence and exhibiting greater robustness to less specialised or manipulated representations, AMD and SAMD may be a useful choice in exploratory settings. However, like all embedding-based approaches, their effectiveness depends on the availability of sufficiently representative corpora that capture the full range of a word’s usages.

Interpretability and lower-dimensional representations. Low-dimensional representations such as definition-based spaces make it possible to relate model behaviour to human-understandable semantic dimensions, enabling targeted qualitative analysis of how and why meanings change. In our experiments, AMD and SAMD remain competitive

under such representations and often outperform APD applied to full embeddings. AMD appears particularly suitable for deeper analysis due to its directional decomposition, which enables fine-grained inspection of asymmetric mismatches between periods via its directional components. In contrast, while SAMD yields strong quantitative performance, its one-to-one matching formulation does not directly support directional interpretability.

These findings align with efforts to move beyond black-box approaches to semantic change detection. Interpretable semantic representations that combine corpus-derived embeddings with structured lexical evidence can provide richer frameworks for understanding semantic change, as explored in linked-data approaches for multilingual diachronic analysis (Armaselu et al., 2024). Other examples include creating interpretable transformations of embedding spaces into structured semantic dimensions, facilitating both analysis and robustness (Cassotti et al., 2024; Aida and Bollegala, 2025). Unlike methods that induce latent sense labels or rely on static prototypes, definition-based projections offer fixed, usage-independent semantic dimensions that can be paired with usage-level metrics to support transparent analysis. More broadly, semantic change metrics that remain robust under dimensionality reduction enable LSCD approaches that balance interpretability with quantitative performance.

Outlook. Overall, our results argue against a one-size-fits-all approach to semantic change metrics. This conclusion aligns with recent large-scale evaluations demonstrating that LSCD methods exhibit highly variable performance depending on the type of semantic change, representation quality, and evaluation conditions (Baes et al., 2025; Periti and Tahmasebi, 2024).

Prototype-based measures, global usage-level metrics, and local correspondence-based metrics may therefore behave differently across research settings, such as low-resource languages, corpora from varying time periods, or specialised domains where embedding quality is weaker. AMD and SAMD should be viewed not as replacements for existing measures, but as complementary additions that

expand the methodological toolkit for contextualised embedding-based LSCD.

6 Conclusion

In this paper, we introduced Average Minimum Distance (AMD) and its Symmetric variant, SAMD, two usage-level metrics for lexical semantic change detection. These metrics are grounded in the simple and intuitive idea that semantic change can be characterised by how well individual usages in one time period can be matched to semantically similar usages in another.

Across a wide range of experiments, we showed that these usage correspondence-based metrics are more robust to both representational quality and encoder choice. When paired with definition-based spaces, AMD improves in performance, demonstrating that interpretability and effectiveness need not be mutually exclusive. This opens up new possibilities for analysing semantic change in controlled and human-understandable spaces.

Overall, our findings show that AMD and SAMD should be considered alongside existing measures as a robust and flexible option for contextualised embedding-based semantic change detection.

Limitations

While our experiments show that APD and PRT degrade more sharply than AMD and SAMD under dimensionality reduction, and we offer hypothetical explanations for this trend, we do not provide a mechanistic account of the underlying causes. Understanding how different semantic change metrics interact with embedding geometry and variance structure remains an open question.

Although AMD and SAMD are more compatible with reduced and interpretable representation spaces, the metrics themselves yields scalar change scores (or two directional scalars), which are not directly interpretable on their own.

Our robustness analysis is limited to different types of dimensionality reduction. We do not explore the behaviour of these metrics under other forms of embedding manipulation, such as noise injection, domain-adaptive fine-tuning, or alternative normalization and align-

ment procedures.

All definitions used in this work are generated using a single large language model (Gemini 2.5 Pro) with a fixed prompting strategy. While this ensures consistency across languages and experiments, variation in the quality, granularity, or coverage of the generated definitions may influence the resulting definition-based spaces and, in turn, the performance of AMD when paired with them. We do not evaluate the sensitivity of our results to alternative LLMs, prompting strategies, or numbers of generated definitions, and future work could explore how such factors affect the stability and interpretability of definition-based semantic change analysis.

While we evaluate AMD across multiple languages and encoders, our analysis is restricted to existing LSCD datasets with curated target words and gold change scores. The extent to which AMD facilitates lexical semantic change discovery, or interpretable analysis remains to be investigated.

7 Acknowledgments

We are grateful to Giacomo De Luca for his insightful feedback and suggestions.

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

- Taichi Aida and Danushka Bollegala. 2025. [SCDTour: Embedding axis ordering and merging for interpretable semantic change detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14775–14785, Suzhou, China. Association for Computational Linguistics.
- Florentina Armasele, Barbara McGillivray, Claudia Liebeskind, Paola Marongiu, Gintarė Valūnaitė Oleškevičienė, Elena Apostol, and Ciprian Truică. 2024. [Multilingual word embedding and linguistic linked open data for tracing semantic change](#). *Rasprave Instituta za hrvatski jezik*, 50(2):219–257.
- Naomi Baes, Raphael Merx, Nick Haslam, Ekaterina Vylomova, and Haim Dubossarsky. 2025. [LSC-eval: A general framework to evaluate methods for assessing dimensions of lexical semantic change using LLM-generated synthetic data](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10905–10939, Vienna, Austria. Association for Computational Linguistics.
- David Bamman and Patrick J Burns. 2020. [Latin bert: A contextual language model for classical philology](#). *arXiv preprint arXiv:2009.10053*.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. [Spanish pre-trained bert model and evaluation data](#). *arXiv preprint arXiv:2308.02976*.
- Pierluigi Cassotti, Stefano De Pascale, and Nina Tahmasebi. 2024. [Using synchronic definitions and semantic relations to classify semantic change types](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4539–4553, Bangkok, Thailand. Association for Computational Linguistics.
- Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. [XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic change](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. [German’s next language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Jing Chen, Emmanuele Chersoni, and Chu-ren Huang. 2022. [Lexicon of changes: Towards the evaluation of diachronic semantic shift in Chinese](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 113–118, Dublin, Ireland. Association for Computational Linguistics.
- Hyung Won Chung, Thibault FÉvry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. [Re-thinking embedding coupling in pre-trained language models](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *arXiv preprint arXiv:2507.06261*.

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. [Revisiting pre-trained models for Chinese natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.
- Georgiana Dinu and Marco Baroni. 2014. [Improving zero-shot learning by mitigating the hubness problem](#). *CoRR*, abs/1412.6568.
- Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. [Time-out: Temporal referencing for robust modeling of lexical semantic change](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. [Definition generation for lexical semantic change detection](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724.
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. [Analysing lexical semantic change with contextualised word representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
- Roksana Goworek and Haim Dubossarsky. 2024. [Toward sentiment aware semantic change analysis](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 350–357, St. Julian’s, Malta. Association for Computational Linguistics.
- Roksana Goworek and Haim Dubossarsky. 2025. [Multilinguality does not make sense: Investigating factors behind zero-shot cross-lingual transfer in sense-aware tasks](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35004–35029, Suzhou, China. Association for Computational Linguistics.
- Roksana Goworek, Harpal Singh Karicut, Hamza Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Satkirti Purighella, Paridhi Gupta, Muhinyia Ndegwa, Bao Khanh Tran, and Haim Dubossarsky. 2025. [SenWiCh: Sense-annotation of low-resource languages for WiC using hybrid methods](#). In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 61–74, Vienna, Austria. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. [Cultural shift or linguistic drift? comparing two computational measures of semantic change](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. [Diachronic word embeddings reveal statistical laws of semantic change](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. [Temporal analysis of language through neural language models](#). In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.
- Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfeldt. 2021. [Operationalizing a national digital library: The case for a Norwegian transformer model](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. [Lexical semantic change discovery](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998.
- Andrey Kutuzov and Mario Giulianelli. 2020a. [Uio-uva at semeval-2020 task 1: Contextualised embeddings for lexical semantic change detection](#). In *SemEval@COLING 2020*.
- Andrey Kutuzov and Mario Giulianelli. 2020b. [UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval)*, pages 126–134.
- Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022a.

- NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.
- Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022b. Contextualized embeddings for semantic change detection: Lessons learned. *Northern European Journal of Language Technology*, 8.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pre-training approach. *CoRR*, abs/1907.11692.
- Martin Malmsten, Love Börjesson, and Chris Hafenden. 2020. Playing with words at the national library of sweden—making a swedish bert. *arXiv preprint arXiv:2007.01658*.
- Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.
- Matej Martinc, Petra Kralj Novak, and Senja Polak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *LREC 2020*.
- Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.
- Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 11:2487–2531.
- Dominik Schlechtweg, Anna HäTTY, Marco del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 732–746.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020a. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020b. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval)*, pages 1–23.
- Dominik Schlechtweg, Sachin Yadav, and Nikolay Arefyev. 2024. The lscd benchmark: A testbed for diachronic word meaning tasks. *arXiv preprint arXiv:2404.00176*.
- Nina Tahmasebi and Haim Dubossarsky. 2023. Computational modeling of semantic change. *arXiv preprint arXiv:2304.06337*.
- Khonzoda Umarova, Lillian Lee, and Laerdon Kim. 2025. Current semantic-change quantification methods struggle with discovery in the wild. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35342–35355, Suzhou, China. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. Lscdiscovery: A shared task on semantic change discovery and detection in spanish. In *Proceedings of the International Workshop on Computational Approaches to Historical Language Change (LChange’ 22)*.

A Number of Generated Definitions

Language	# Words	Defs / Word
English	37	11.00 ± 5.36
German	48	3.75 ± 1.70
Swedish	31	4.26 ± 1.48
Latin	40	6.20 ± 1.14
Spanish	100	6.84 ± 4.14
Chinese	40	3.85 ± 1.71
Norwegian_1	40	4.75 ± 1.88
Norwegian_2	40	5.00 ± 1.99

Table 1: Number of target words and average number of generated definitions per word (mean ± standard deviation) for each language.

B Definition Generation Prompts

C Qualitative Analysis and Interpretability with AMD

Beyond quantitative evaluation, AMD enables a more fine-grained and interpretable analysis of semantic change through its directional

Language	Prompt
English	Write all dictionary definitions of {WORD} in English. One sense per line. Output only the definitions. Do not add any text before or after the definitions.
German	Schreibe alle Wörterbuchdefinitionen von {WORD} auf Deutsch. Eine Bedeutung pro Zeile. Gib nur die Definitionen aus. Füge keinen Text vor oder nach den Definitionen hinzu.
Swedish	Skriv alla ordboksdefinitioner av {WORD} på svenska. En betydelse per rad. Skriv endast definitionerna. Lägg inte till någon text före eller efter definitionerna.
Latin	Scribe omnes definitiones dictionarii verbi {WORD} Latine. Una significatio per lineam. Redde tantum definitiones. Ne quidquam addas ante aut post definitiones.
Spanish	Escribe todas las definiciones de diccionario de {WORD} en español. Un significado por línea. Devuelve solo las definiciones. No añadas ningún texto antes ni después de las definiciones.
Chinese	用中文写出“{WORD}”的所有词典释义。 每行一个义项。只输出释义，不要在释义前后添加任何文字。
Norwegian	Skriv alle ordbokdefinisjoner av {WORD} på norsk. Én betydning per linje. Svar kun med definisjonene. Ikke legg til tekst før eller etter definisjonene.

Table 2: Language-specific prompts used for dictionary definition generation.

components. Recall that the directional scores $\text{AMD}(A \rightarrow B)$ and $\text{AMD}(B \rightarrow A)$ capture asymmetric mismatches between periods. High $\text{AMD}(A \rightarrow B)$ indicates usages in the earlier corpus that cannot be well matched in the later corpus, suggesting sense narrowing or disappearance, while high $\text{AMD}(B \rightarrow A)$ highlights usages emerging in the later corpus, corresponding to sense broadening or innovation. Words exhibiting high values in both directions reflect overall shifts in meaning.

To illustrate this behaviour, we identify words with the largest imbalance between directional AMD scores (highest differences between them), highlighting cases of predominantly unidirectional change. Across both full embedding space and definition-based space (produced with XL-LEXEME), we observe patterns of narrowing (e.g., *plane*, *head*) and broadening/emergence of new senses (e.g., *record*, *graft*). While individual rankings vary slightly, the directional decomposition consistently surfaces words whose usage distributions change asymmetrically across time.

The definition-based space further enables a more explicit interpretation of these changes. For a given target word, we apply linear discriminant analysis (LDA) to its usage embeddings projected into definition-based space, identifying the direction that best separates usages from the two time periods. This yields

a weighted combination of definition dimensions, where positive weights indicate stronger association with the later corpus and negative weights with the earlier corpus. Inspecting the most discriminative dimensions provides a direct, human-interpretable signal as to which aspects of meaning are more or less represented between the two corpora (see Table 3).

For example, for *plane*, LDA highlights a shift from tool-related meanings to air-travel-related usages, while for *record*, the most discriminative dimensions correspond to a move from instrumental readings to official and documentary uses.

These examples illustrate how AMD, particularly when paired with interpretable lower-dimensional representations, supports targeted qualitative inspection of semantic change. Rather than treating semantic shift as a single scalar quantity, this framework enables distinguishing between different change patterns—such as sense emergence, disappearance, or redistribution—and relating them to interpretable semantic dimensions, enabling fine-grained inspection of semantic change beyond what is possible with global distance metrics based on full embeddings.

Word	Earlier-associated definition	Later-associated definition
<i>plane</i>	A tool consisting of a block with a projecting blade, for shaping and smoothing a wooden surface.	To travel in an airplane.
<i>record</i>	To indicate or show (a reading) on a measuring instrument.	To state for the purpose of being set down in an official account.

Table 3: Most discriminative definition dimensions identified by LDA in definition-based space for selected words. Negative weights are associated with the earlier corpus and positive weights with the later corpus.

D Hubness Analysis of Embedding Spaces

Nearest-neighbour-based similarity measures in high-dimensional spaces are known to be affected by the *hubness phenomenon*, whereby a small number of points appear disproportionately often as nearest neighbours of other points due to distance concentration and anisotropic embedding geometry. In the context of lexical semantic change detection, hubness may bias metrics such as Average Minimum Distance (AMD) if a small number of transitional or atypical historical usages act as hubs for many modern usages. For instance, a historical usage with unusually modern-like contextual properties could become the nearest neighbour for a large fraction of modern embeddings, leading AMD to underestimate broader distributional change.

To quantify hubness across embedding spaces, we compute three symmetric hubness statistics derived from nearest-neighbour occurrence distributions: (i) **dominant nearest-neighbour share**, the proportion of embeddings whose nearest neighbour is the single most frequent hub; (ii) **unused nearest-neighbour share**, the proportion of embeddings that are never selected as nearest neighbours (anti-hubs); and (iii) **average nearest-neighbour load**, the average number of embeddings assigned to each embedding that serves as a nearest neighbour at least once. Each measure is computed in both temporal directions and averaged to obtain symmetric scores.

Overall, hubness effects are moderate and comparable across spaces, and do not exhibit pathological collapse that would explain AMD’s performance gains. Dominant nearest-neighbour share remains around 2–27, indicating that only a minority of embeddings collapse onto a single hub, while unused nearest-

neighbour share lies between 5 and 6, reflecting selective but non-pathological neighbourhood structure typical of high-dimensional embeddings. Importantly, the definition-projected space does not exhibit increased hub dominance relative to the original embedding space, and in fact slightly reduces both dominant and unused nearest-neighbour shares. The definition space shows a higher average nearest-neighbour load with greater variance, reflecting stronger local clustering around semantic prototypes induced by projection rather than pathological hub collapse.

To further assess robustness to hubness, we implement a symmetric variant of AMD based on greedy one-to-one matching between embeddings across corpora. Unlike standard AMD, which relies on independent nearest-neighbour assignments, this symmetric formulation constructs approximate optimal matchings and is therefore substantially less susceptible to hubness effects. Empirically, the symmetric matching AMD consistently outperforms APD and PRT and closely matches or improves upon standard AMD across spaces. This indicates that performance gains are not driven by hubness artefacts, but reflect improved semantic alignment between temporal corpora.

Overall, these results suggest that hubness does not account for the effectiveness of nearest-neighbour-based semantic change metrics in our setting, and that definition-based projection does not amplify hubness effects.

Metric	Full	Definition	PCA	Random
Dominant NN share	.268 (.159)	.259 (.167)	.243 (.134)	.197 (.139)
Unused NN share	.616 (.069)	.578 (.100)	.583 (.090)	.497 (.049)
Average NN load	2.882 (.618)	3.175 (1.521)	2.923 (.941)	2.330 (.648)

Table 4: Mean (standard deviation) of hubness measures across spaces, aggregated over languages and encoders.

From *sunblock* to *softblock*: Analyzing the correlates of neology in published writing and on social media

Maria Ryskina^{1‡}, Matthew R. Gormley², Kyle Mahowald³, David R. Mortensen², Taylor Berg-Kirkpatrick⁴, Vivek Kulkarni

¹Vector Institute for Artificial Intelligence, ²Carnegie Mellon University,

³The University of Texas at Austin, ⁴University of California, San Diego

[‡]Work done partly at Carnegie Mellon University

Correspondence: mryskina@alumni.cmu.edu

Abstract

Living languages are shaped by a host of conflicting internal and external evolutionary pressures. While some of these pressures are universal across languages and cultures, others differ depending on the social and conversational context: language use in newspapers is subject to very different constraints than language use on social media. Prior distributional semantic work on English word emergence (*neology*) identified two factors correlated with creation of new words by analyzing a corpus consisting primarily of historical published texts (Ryskina et al., 2020b). Extending this methodology to contextual embeddings in addition to static ones and applying it to a new corpus of Twitter posts, we show that the same findings hold for both domains, though the topic popularity growth factor may contribute less to neology on Twitter than in published writing. We hypothesize that this difference can be explained by the two domains favouring different neologism formation mechanisms.¹

1 Introduction

Hypotheses about the mechanisms of language evolution have traditionally been tested by comparing linguistic data over long periods of time (Aitchison, 2001). Until the end of the last century, the majority of the textual sources preserved across many decades have been books, periodicals, or other forms of published writing. Though analyses of such texts can yield important discoveries, they do not represent the entirety of their time’s everyday language use. Institutions like publishers and news agencies tend to favour standardization (Milroy and Milroy, 2012) and adopt linguistic innovation less readily than individuals: by the time a word appears in print, it is likely already widely

used.² Historical print media data is also unlikely to reflect the language of underrepresented and marginalized communities—a fertile ground for linguistic creativity (Montgomery, 2008; Lefkowitz and Hedgcock, 2017).

The Internet has opened up new avenues for observing and studying language change (McCulloch, 2020). Chats, blogs, and social media contain unprecedented amounts of linguistic data and represent a greater diversity of voices and styles. This richness makes online communication especially well-suited for population-level analysis of the evolutionary forces (Bower, 2019) that drive innovation, such as social prestige, cognitive economy, or communicative need—the language users’ collective desire to express certain meanings.

We have previously tested two hypotheses formalizing the role of communicative need in word emergence (*neology*; Ryskina et al., 2020b). Analyzing a historical corpus primarily comprised of books and other publications, we found that new words are created both to give names to new concepts in domains of growing cultural importance and to fill in gaps in the word distribution over the space of possible meanings, with the former factor being more significant. In this paper, we ask: will an analysis of the same factors in social media neology yield similar findings? And if there are differences, are they simply due to the limitations of historical published corpora (e.g., sparsity), or are the underlying pressures of word emergence different between the two domains?

We collect a corpus of 260M English-language tweets (§4.2) and extend our original methodology (Ryskina et al., 2020b) to study the correlates of neology both on social media and in published writing. Experimenting with different operationalizations (e.g., using both static and contextual word

¹Code, word lists, and tweet IDs can be found at [https://github.com/ryskina/twitter-neology/](https://github.com/ryskina/twitter-neology)

²Rogers and Schoemaker (1971, p. 255) conclude that adoption of innovation in general is more likely to spread through interpersonal channels than through mass media.

embeddings; §5.2) of our two hypotheses (§3), we robustly reproduce our earlier findings for a corpus of published writing (support for both hypotheses; §6.1). We then perform the same analysis for Twitter and observe similar trends (§6.2), though frequency growth of the neighbouring words may play a less important role there.

2 Related work

NLP studies of neology and word decline Several computational studies have focused on the mechanisms and driving factors of word emergence (and loss) in languages, building on cognitive and linguistic models (Xu et al., 2019, 2023; Karjus et al., 2020, 2021). One line of work uses this knowledge to trace the etymological origins of neologisms (Kulkarni and Wang, 2018; Wu and Yarowsky, 2020), especially lexical blends (Cook and Stevenson, 2010; Pinter et al., 2020). Other studies model the process as a whole, predicting what makes words or senses likely to emerge (Hofmann et al., 2020; Sun and Xu, 2022), persist (Stewart and Eisenstein, 2018), or decline (Francis et al., 2021; Kali et al., 2024). Most relevant to the current study is our prior work (Ryskina et al., 2020b) which tests two hypotheses about intra- and extralinguistic pressures that facilitate word emergence (§3); this study builds primarily on the experimental methodology introduced there.

Processing of novel words remains a significant challenge in NLP applications. Neologisms, along with synthetic nonce words, are used to test model generalization in domains like translation (Zheng et al., 2024; Lerner and Yvon, 2025), morphology (Weissweiler et al., 2023; Mortensen et al., 2024), and definition generation (Malkin et al., 2021).

Diachronic analysis of published writing Large time-stamped corpora of texts lend themselves to comparative analyses of historical language change. Such corpora typically span decades or centuries and include literature (Michel et al., 2011; Bouma et al., 2020), news articles (Popescu and Strapparava, 2015; Basile et al., 2020), or both (Onelli et al., 2006; Davies, 2009, 2012).

In the computational linguistics community, such corpora have most often been used for tracking the evolution of word meanings (Kutuzov et al., 2018; Tahmasebi et al., 2021), typically quantified by measuring changes in word embeddings across time. Earlier approaches included learning static embeddings from different time periods

(Kulkarni et al., 2015; Hamilton et al., 2016) or building temporally-aware dynamic embeddings (Del Tredici et al., 2016; Rudolph and Blei, 2018), and the rise of pretrained language models made contextual embeddings more popular (e.g., Giulianelli et al., 2020; Martinc et al., 2020; Montariol et al., 2021; Periti and Tahmasebi, 2024; see Periti and Montanelli, 2024 for a survey).

Diachronic book corpora are also used for evaluating text-dating models (Kulkarni et al., 2018; Ren et al., 2023), analyzing change in syntactic phenomena (Krielke et al., 2022; Miletic and Schulte im Walde, 2025), and applications in digital humanities (Ruiz Fabo et al., 2017; Klausner and Vogel, 2018; Haider and Eger, 2019).

Diachronic analysis of social media Social media is an exceptionally rich domain for studying language change. Users adopt novel words and spellings to express their creativity, signal community membership, or avoid moderation in scenarios ranging from protecting marginalized users from censorship (Ungless et al., 2025) to disguising dog whistles (Kruk et al., 2024; Sasse et al., 2025).

Most existing quantitative work on language change in social media focuses on tracking the life cycle of emergent words (Danescu-Niculescu-Mizil et al., 2013; Würschinger et al., 2016; Stewart and Eisenstein, 2018; Würschinger, 2021) and the social, demographic, and geographical factors that affect it (Eisenstein et al., 2014; Grieve et al., 2018; Ananthasubramaniam et al., 2024, 2025). To our knowledge, ours is the first study of the semantic factors driving word emergence on social media.

3 Question and hypotheses

We define ‘neologism’ broadly as a novel form–meaning pair that at some point enters more general use (as reflected by a substantial increase in the form’s usage frequency; Ryskina et al., 2020b). This includes new coinages (*yeet*), combinations of known elements (*tiktoker*, *cryptocurrency*), or existing words paired with new senses (*transformer*). In particular, we are interested in *what kinds of meanings* such neologisms are likely to express.

We reuse the supply-driven and demand-driven neology hypotheses introduced in our prior work (Ryskina et al., 2020b). These competing hypotheses rely on the concept of a semantic space—a manifold of meanings where certain points correspond to words—and posit where in such a space neologisms are likely to emerge. We operationalize

the hypotheses under the distributional semantics paradigm (Lenci and Sahlgren, 2023), using word embedding spaces informed by the co-occurrence statistics within our corpora as a proxy for the underlying semantic space.

Supply hypothesis This hypothesis suggests that neologisms are more likely to emerge in *sparser areas of the semantic space*. It is derived from Bréal’s (1904) law of differentiation, which states that the semantic space tends towards uniformity; if two existing words are too close in meaning, they will either diverge or one of them will fall out of use. By extension, we posit that the existence of gaps in the semantic space could create pressure on the language to repair uniformity by populating them with neologisms.

Demand hypothesis The second hypothesis states that neologisms are more likely to emerge in *semantic neighbourhoods of growing popularity*. Growing popularity of a certain semantic cluster—i.e., growing frequency of use for the words that make it up—can be viewed as a reflection of the increasing importance of the corresponding domain of discourse. Rapidly developing domains such as technology might produce novel concepts faster, and the need for words to refer to these new concepts could also be a driving factor of neology.

4 Data

4.1 Published writing corpora

As in our 2020b work, we consider two collections of American English texts from non-overlapping time periods: HISTORICAL (1800–1989) and MODERN (1990–2012). We will denote these two datasets as $\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$ and $\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$ respectively. Both are drawn from existing multi-genre diachronic corpora: COHA (Davies, 2012) for $\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$ and COCA (Davies, 2009) for $\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$. In both corpora, all texts are already annotated for genre and time period (decade or year for COHA and COCA respectively), preprocessed, and tokenized; we do not perform any additional preprocessing besides sentence-splitting. In order to be able to estimate the frequency trends in $\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$, we split this dataset into decades using the temporal labels provided with the COHA corpus. We restrict both datasets to published materials only: fiction, non-fiction, academic articles, and popular magazine and newspaper articles. The

statistics for both $\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$ and $\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$ are reported in Table 1.

4.2 Twitter data collection

We collect a corpus of $\sim 260\text{M}$ tweets using the Twitter API,³ spanning the period from January 2007 to December 2021. Aiming for uniform coverage, we randomly sample 100K unique tweets per day; if the total number of tweets available for a given day is $< 100\text{K}$ (e.g., in the early days of the platform’s existence), we use the entire set. We restricted the tweets to be in English only and excluded retweets or tweets from bots. Since there were much fewer users on the platform in the initial years after its launch in 2007, the tweet distribution in our corpus is skewed: the number of tweets per year grows from $\sim 80\text{K}$ in 2007–2008 to 7M in 2009 to 18–24M in 2010–2021. We discard tweets that contain only URLs or other special tokens (0.3% of tweets total; details in Appendix B.1).

We split the collected data into the HISTORICAL (representing the “baseline” distribution of word use) and MODERN (in which the neologisms appear) subsets, denoted $\mathcal{D}_{\text{HISTORICAL}}^{\text{Twt}}$ and $\mathcal{D}_{\text{MODERN}}^{\text{Twt}}$. All tweets are labelled by year; we allocate tweets from 2007–2010 to the HISTORICAL subset and tweets from 2011–2021 to the MODERN subset. The statistics for both $\mathcal{D}_{\text{HISTORICAL}}^{\text{Twt}}$ and $\mathcal{D}_{\text{MODERN}}^{\text{Twt}}$ are reported in Table 1.

5 Methodology

We build upon the methodology proposed in our previous study (Ryskina et al., 2020b), with several modifications detailed in this section and in Appendix A. First, neologisms are identified automatically and filtered (§5.1). Second, each of them is paired with a non-neologism control word that is similar to the neologism in several aspects (§5.3). Finally, we compare the distribution of the statistics of interest across the neighbours of the neologisms and the control words (§5.4).

5.1 Neologism selection

Published writing We reuse the list of neologisms extracted from the COCA-based $\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$ data used in our 2020b study. It includes 1000 nouns that appear substantially more frequently in the $\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$ split than in the $\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$ one.⁴

³Last accessed in January 2023.

⁴Note that the $\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$ corpus from the 2020b study is a superset of the one in this work: it included speech transcripts, excluded from this study due to focus on published writing.

	Texts	Sentences	Tokens	Types	Avg. text length	Avg. sentence length
$\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$	93K	21M	385M	1.8M	4,140 tokens	18 tokens
$\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$	151K	24M	406M	2.5M	2,704 tokens	17 tokens
$\mathcal{D}_{\text{HISTORICAL}}^{\text{Tw}}$	31M	48M	409M	3.0M	13 tokens	9 tokens
$\mathcal{D}_{\text{MODERN}}^{\text{Tw}}$	229M	331M	3,206M	9.6M	14 tokens	10 tokens

Table 1: Corpus statistics for the published writing data (rows 1–2) and the Twitter data (rows 3–4). The data for each domain is split into a HISTORICAL and a MODERN subset as described in §4. For the published writing domain, ‘text’ = book, story, or article; for Twitter, ‘text’ = tweet. All texts are split into sentences using the NLTK toolkit (Bird et al., 2009). $\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$ and $\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$ data is already split at the token level (Davies, 2009, 2012), and we use NLTK’s Twitter-specific tokenizer for tweets. The ‘Tokens’ column shows the total number of tokens in each subcorpus, while the ‘Types’ column shows the number of unique tokens in each subcorpus.

Twitter Since social media data exhibits much greater lexical diversity than published works, simply comparing frequencies in $\mathcal{D}_{\text{HISTORICAL}}^{\text{Tw}}$ and $\mathcal{D}_{\text{MODERN}}^{\text{Tw}}$ results in a noisy set of candidates. Instead, we follow the method of Kulkarni et al. (2018), which for a given word estimates the year when it came into popular usage. Given a diachronic corpus spanning timesteps $\{1, \dots, T\}$ (here corresponding to years $y \in \{2007, \dots, 2021\}$, i.e., $t = y - 2006$) and a word w , they find the timestep t when the cumulative usage of w first exceeds a specific percentage of its total cumulative usage through the entire corpus:

$$t^* = \arg \min_t \left[\sum_{i=1}^t c_i(w) > \alpha \cdot \sum_{j=1}^T c_j(w) \right],$$

where $c_t(w)$ represents how many times the word w was used in the timestep t . Words which came into popular usage during the MODERN period ($y^* = t^* + 2006 \geq 2011$) are selected as potential neologisms. Empirically we set $\alpha = 1/300$.

We further refine the resulting list using part-of-speech tagging (e.g., to remove proper nouns) and frequency-based heuristics (to remove rare variants or filter out auto-generated tweets); see Appendix B.2 for details. Unlike our 2020b study, here we do not restrict the neologisms to nouns only. After this step, 938 potential neologisms are left.

Manual filtering For a stricter analysis, we use dictionaries and other resources to filter out words that would not be considered neologisms per our definition (§3). For the published writing domain, we remove candidate neologisms that have been in use prior to 1900 and have not gained new senses since then. For Twitter, we manually classify all candidate neologisms to remove all proper names,

foreign words, typos and errors, and words that have been in use before 2000 and have not gained new senses since. The details of the filtering procedure are described in Appendix B.3. The results reported in the following sections are based on this filtered set of neologisms; the same for the non-filtered word list can be found in Appendix D.

5.2 Embeddings

One common approach in modelling semantic shift is learning separate sets of static word embeddings from subcorpora that represent different time periods and then projecting them into a common space for comparison (Tahmasebi et al., 2021). Recently, contextual embeddings extracted from BERT (Devlin et al., 2019) or related models have replaced static embeddings as the tool of choice (Periti and Tahmasebi, 2024). We experiment with both types to test our word emergence hypotheses.

Static embeddings We train separate Word2Vec SkipGram embeddings (Mikolov et al., 2013; window size = 5, dimension = 300) on the HISTORICAL and the MODERN subcorpora. In prior work on language change, static embedding spaces from distinct time periods are often projected into the same axes via a linear transformation, which seeks to align a set of anchor points between the two spaces (Kulkarni et al., 2015; Zhang et al., 2015, 2016; Hamilton et al., 2016). Specifically, we use an orthogonal Procrustes transformation (following the setup of Hamilton et al., 2016) with all words that exist in the vocabularies of both models used as anchors. The alignment step is necessary for finding the neighbourhoods in the HISTORICAL space where neologisms eventually appear: as most neologisms are not in the vocabulary of the HISTORICAL Word2Vec model, we approximate their

positions in the HISTORICAL space by projecting their MODERN vectors into the HISTORICAL axes.

We set the minimum word frequency threshold for both the HISTORICAL and the MODERN vocabularies to keep them under 100K words (resulting in $\sim 98\text{K}$ vocabulary words for $\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$, $\sim 99\text{K}$ words for $\mathcal{D}_{\text{MODERN}}^{\text{Pub}}$, and $\sim 100\text{K}$ words for $\mathcal{D}_{\text{HISTORICAL}}^{\text{Twt}}$ and $\mathcal{D}_{\text{MODERN}}^{\text{Twt}}$). These vocabularies, denoted V_H and V_M , are used for the rest of the experiments with both static and contextual embeddings.

Contextual embeddings These embeddings are richer and better for capturing sense variation. However, having many embeddings for the same word makes it difficult to define the “neighbour” relationship between words, which is central for our hypotheses. To avoid this, we average the embeddings to reduce them to their static versions (Bommasani et al., 2020).

We extract the 768-dimensional contextual embeddings from the RoBERTa model, pretrained on a large corpus of English-language books, news articles, Wikipedia articles, and other online content (Liu et al., 2019). During pretraining, 15% of the words in each sentence were randomly masked, and the model was trained to predict the masked words based on their bidirectional context. We download the pretrained 12-layer RoBERTa-Base model from the Hugging Face Hub.⁵

We follow the procedure of Timkey and van Schijndel (2021) for obtaining the embeddings. For each of the ~ 99 – 100K HISTORICAL vocabulary words we sample 250 context sentences from $\mathcal{D}_{\text{HISTORICAL}}^{\{\text{Pub}, \text{Twt}\}}$.⁶ Similarly, for each neologism word we sample 500 context sentences from $\mathcal{D}_{\text{MODERN}}^{\{\text{Pub}, \text{Twt}\}}$. Each word’s embedding in each sampled context (mean-pooled over subword tokens) is extracted from the last layer of RoBERTa-Base. We z-score all obtained vectors to get rid of the “rogue dimensions” that affect word vector similarity (Timkey and van Schijndel, 2021). Finally, we average each word’s embeddings across its sampled contexts to obtain one static, decontextualized vector per word.

5.3 Control set selection

To see how the neighbourhoods of neologisms differ from the neighbourhoods of other similar non-neologism words, we pair each neologism with a

⁵<https://huggingface.co/FacebookAI/roberta-base>

⁶Some vocabulary words have fewer/no contexts because of tokenization mismatches between RoBERTa and the tokenizer used to create the vocabulary (Appendix B.1).

control word (Dubossarsky et al., 2017). As in prior work, we control for word frequency and length (Ryskina et al., 2020b; Francis et al., 2021), and additionally constrain the control to be semantically similar to the neologism. Formally, for each neologism w_n we select a counterpart w_c satisfying the following constraints:

- Frequency ranks of the two words in the corresponding corpora are in the same percentile: $\left| \frac{z_M(w_n)}{|V_M|} - \frac{z_H(w_c)}{|V_H|} \right| \leq 0.01$. Here $1 \leq z_H(\cdot) \leq |V_H|$ and $1 \leq z_M(\cdot) \leq |V_M|$ are ranks of the words in the HISTORICAL and MODERN vocabularies, sorted by frequency;
- The length of the two words is identical up to 3 characters;
- The cosine similarity between the neologism and its control counterpart in the HISTORICAL embedding space is above a certain threshold: $\text{cosine}(v_{w_n}, v_{w_c}) \geq 0.4$. Since we want to keep the neologism–control pairs consistent in all experiments, we only use Word2Vec embeddings for this pairing step. Here v_w denotes a projected MODERN static embedding if w is a neologism or a HISTORICAL static embedding if w is a control word.

We use a maximum bipartite matching algorithm (Hopcroft and Karp, 1973) to pair neologisms and controls, finding matches for 231 of the 459 Twitter neologisms and 557 of the 746 published writing neologisms. After this step, neologisms and controls are only compared as sets; we do not perform any statistical comparison within individual pairs.

Domain	Neologism	Control
Published writing (1810–2012)	e-mail	message
	sunblock	nicotine
	sitcom	cinema
	dysfunction	functional
Tweets (2007–2021)	cringiest	silliest
	softblock	un-followed
	bruhhhhh	niceeeee
	baecation	staycation

Table 2: Example neologism–control word pairs extracted from either corpus.

Example neologism–control pairs for either data domain are shown in Table 2. We find that the

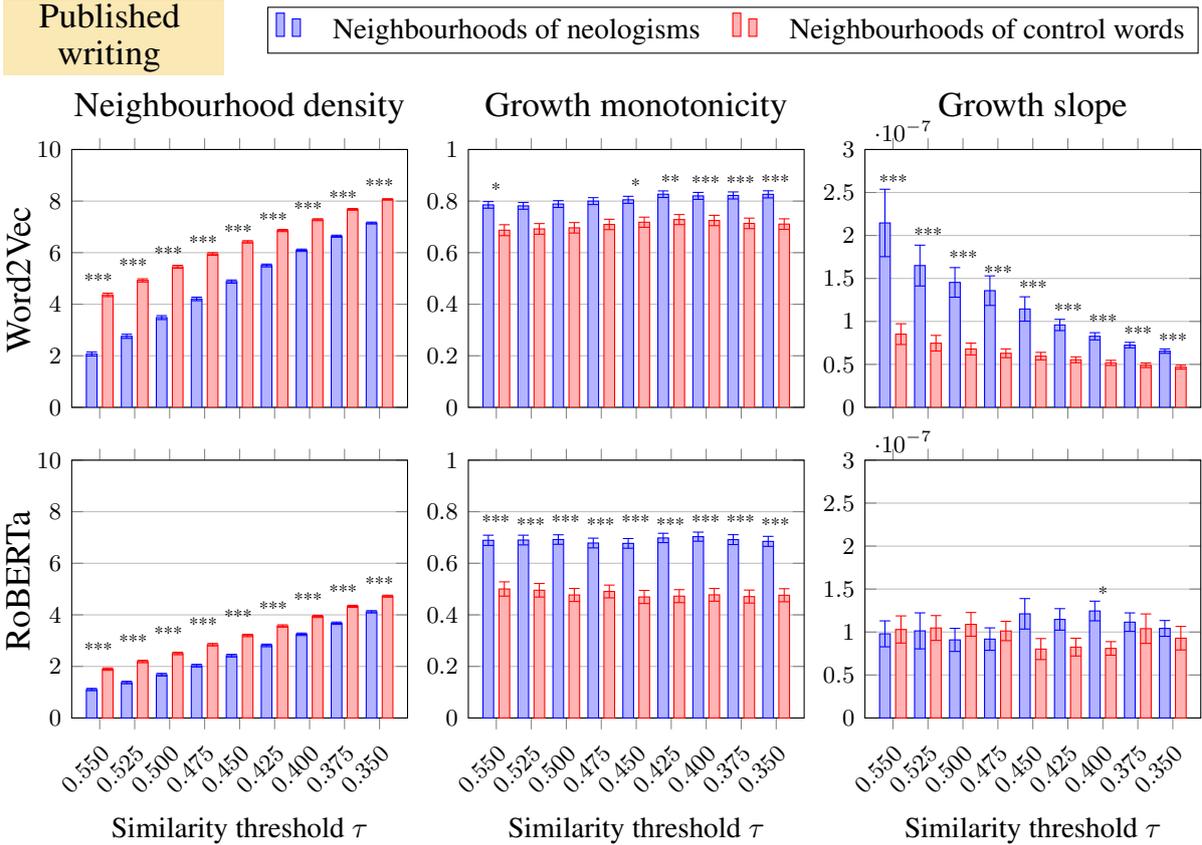


Figure 1: Experimental comparison between the neighbourhoods of neologisms (blue bars) and control words (red bars) in the **published writing** domain. The three plots in each row correspond to three measures: the number of HISTORICAL neighbours a word has (left), how monotonically these neighbours grow in frequency (centre), and the linear regression slope of their growth (right). The x-axis on all plots corresponds to the neighbourhood size (defined by the cosine similarity threshold τ). The top and bottom rows show the results with the static Word2Vec embeddings and the contextual RoBERTa embeddings respectively. Error bars represent standard error over words. The number of asterisks above a pair of bars indicates the statistical significance of their difference per Wilcoxon signed-rank test: *** for $p < 0.001$, ** for $0.001 \leq p < 0.01$, * for $0.01 \leq p < 0.05$, none for $p \geq 0.05$.

cosine similarity constraint often adds an extra semantic or morphological connection to the resulting word pairs. It can manifest as conceptual similarity (e-mail:message), morphological overlap (baecation:staycation), matching part of speech and form (cringiest:silliest), or creative spelling choice (bruhhhhh:niceeeee).

5.4 Experimental setup

We operationalize the supply and demand hypotheses following our prior work (Ryskina et al., 2020b), with minor modifications to make the analysis more robust (detailed comparison in Appendix A). Let $N_\tau(w)$ denote a neighbourhood of a word w , i.e., the set of words in the HISTORICAL embedding space whose cosine similarity to w exceeds a given threshold τ . If w is a neologism, which may not be present in the HISTORICAL vocabulary V_H , we use its MODERN embedding (Procrustes-

projected for Word2Vec or as-is for RoBERTa).

To quantify the supply hypothesis, we measure a function of the neighbourhood density:

$$d(w, \tau) = \log(|N_\tau(w)| + 1). \quad (1)$$

The demand hypothesis measures how the neighbourhood words change in popularity. At each timestep $t \in \{1, \dots, T\}$, we measure how much of the corpus do the words in $N_\tau(w)$ account for:

$$p_\tau(w, t) = \frac{\sum_{u \in N_\tau(w)} c_t(u)}{\sum_{u \in V_H} c_t(u)} \quad (2)$$

We measure the monotonicity of this time series (Spearman correlation ρ with the timesteps):

$$m(w, \tau) = \rho(\{1, \dots, T\}, \{p_\tau(w, t)\}_{t=1}^T) \quad (3)$$

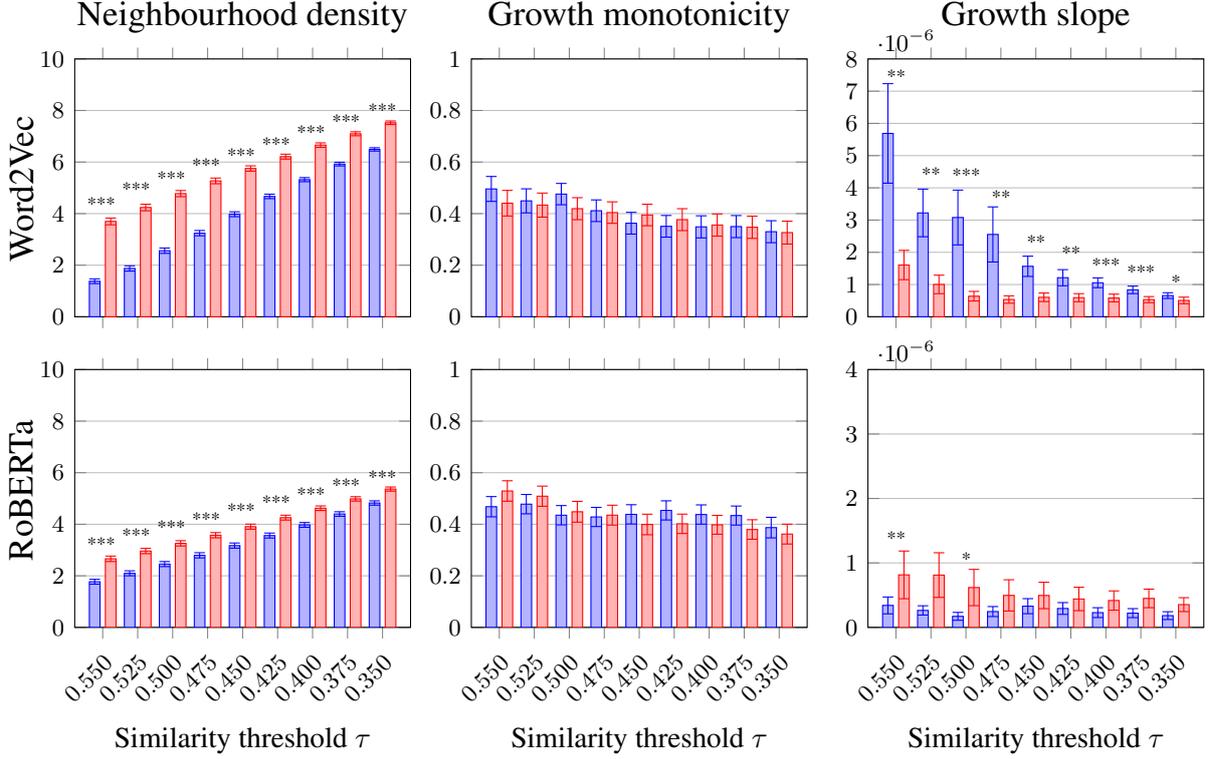


Figure 2: Experimental comparison between the neighbourhoods of neologisms (blue bars) and control words (red bars) in the **Twitter** domain. The three plots in each row correspond to three measures: the number of HISTORICAL neighbours a word has (left), how monotonically these neighbours grow in frequency (centre), and the linear regression slope of their growth (right). The x-axis on all plots corresponds to the neighbourhood size (defined by the cosine similarity threshold τ). The top and bottom rows show the results with the static Word2Vec embeddings and the contextual RoBERTa embeddings respectively. Error bars represent standard error over words. The number of asterisks above a pair of bars indicates the statistical significance of their difference per Wilcoxon signed-rank test: *** for $p < 0.001$, ** for $0.001 \leq p < 0.01$, * for $0.01 \leq p < 0.05$, none for $p \geq 0.05$.

and the linear regression slope of its growth:

$$r(w, \tau) = \frac{1}{d(w, \tau)} \times \frac{\sum_{t=1}^T (t - \bar{t})(p_\tau(w, t) - \bar{p}_\tau(w))}{\sum_{t=1}^T (t - \bar{t})^2}, \quad (4)$$

where $\bar{p}_\tau(w) = \frac{1}{T} [\sum_{t=1}^T p_\tau(w, t)]$ and $\bar{t} = \frac{(1+T)}{2}$.

In the experiments described in the next section, we compute the mean values of these three metrics for the neighbourhoods of neologisms and the neighbourhoods of control words over a range of neighbourhood sizes defined by the threshold τ .

6 Results

6.1 Published writing

Figure 1 shows how the mean neighbourhood density (left) and the mean neighbour frequency

growth rate (middle, right) differ between neologisms and controls in the published writing corpus over a range of neighbourhood sizes. The top row of charts shows the results for the static embeddings experiment, which we perform to ensure that our original findings (Ryskina et al., 2020b) still hold given the changes in the neologism–control pairing criteria (§5.3) and the operationalization of the hypotheses (§5.4). We successfully reproduce the original results, finding support for both the supply and the demand hypotheses: neologisms have fewer neighbours than the control words (Figure 1, top left) but their neighbours grow in frequency faster (Figure 1, top centre & right).

The bottom row of charts shows the results with the RoBERTa embeddings: we confirm that the trends remain the same when we perform the experiment using contextual embeddings. However,

Neologism formation mechanism	Published writing (1810–2012) % neologisms (examples)	Twitter (2007–2021) % neologisms (examples)
Abbreviation	6% (blog, condo, postdoc)	20% (bae, afab, incel)
Blend	4% (pixel, camcorder)	14% (presstitutes, weeksary)
Borrowing	14% (falafel, video)	11% (mukbang, dololo)
Compound	45% (laptop, cyberpunk)	19% (deepfake, headcanon)
POS conversion	7% (startup, aerobics)	8% (snapchatting, mutuals)
Derivation	29% (wellness, facilitator)	20% (anti-mask, unalive)
Sense	4% (analog, firewall)	4% (catfishing)
Spelling	<1% (byte, gangsta)	13% (bæ, 5ever, sksksk)

Table 3: Percentage of neologisms in each domain by their formation mechanism. Our manual categorization draws on the Oxford English Dictionary where available, and allows for the same word to fall into multiple categories. Category definitions (details in §C): abbreviation = clipping, shortening, acronym; blend = combining parts of existing words; borrowing = loanwords from other languages; compound = combining full existing words; POS conversion = change in part of speech; derivation = morphological derivation, e.g., affixation; sense = widely used existing words acquiring new senses; spelling = any creative spellings, including alterations of existing words.

the significance of the findings for the demand hypothesis is less conclusive: the difference in neighbour frequency growth between neologism and control neighbourhoods is less pronounced for the regression slope metric (Figure 1, bottom right) than for the monotonicity metric (Figure 1, bottom centre), unlike the static embedding experiment where both metrics show a significant difference for most neighbourhood sizes.

6.2 Twitter

The results for our Twitter corpus are presented in Figure 2. As in the published writing corpus, we find significant evidence for the supply hypothesis (left column), suggesting that neology on social media is also partly driven by the pressure to fill in gaps in the lexicon. However, the evidence for the demand hypothesis is weaker: the frequency growth monotonicity metric (middle column) does not show a significant difference between neologisms and controls, and the frequency growth slope (right column) is higher for neighbours of neologism than for neighbours of control words only with Word2Vec embeddings (top right). This suggests that the frequency growth is less important in Twitter neology, though we discuss some alternative explanations in the following section.

7 Discussion

Domain and corpus differences The informal and participatory nature of social media encourages creativity (Peppler, 2013). With language in particular, additional constraints like automated

moderation (Ungless et al., 2025) or typing and text rendering issues (Ryskina et al., 2020a) further incentivize users to experiment with spelling and word creation. These domain-specific pressures are reflected in how social media neologisms are formed and in what neighbourhoods they emerge.

We categorize all verified neologisms by the mechanism of their formation, using dictionaries and other resources (Table 3; details in Appendix C). Vast majority of the neologisms in published writing are created by recombining existing morphemes, either through compounding (e.g., *airfare*) or morphological derivation (e.g., *interconnectedness*). As shown in Table 3, Twitter neologisms are much more diverse and creative. Blends and portmanteaus are more prominent, especially in fandom culture (e.g., *Bettie + Archie* → *Barchie*). Expressions of orthographic creativity are common, including phonetic respellings of existing words and phrases (e.g., *on that* → *onnat*), expressing intonation or emotion (e.g., *stop* → *stahp*, *lit* → *littttt*), puns and wordplay (e.g., *forever* → *4ever* → *5ever*, *democrats* → *demonrats*), and even completely novel coinages (e.g., *sksksk* to express laughter). Twitter’s character limit also encourages abbreviation, such as clipping (*cryptocurrency* → *crypto*), shortening (*season* → *szn*), and acronyms (*one of my followers* → *oomf*). Finally, our published writing corpus focuses on American English and is mostly representative of its “standard” variety (Milroy and Milroy, 2012), while the Twitter data features African American English (known to be one of the main sources of lexical innovation on

the platform; Grieve et al., 2018), other World Englishes, and code-switching with other languages (Doğruöz et al., 2021). Although the overall share of neologisms of foreign origin (see ‘Borrowing’ in Table 3) is lower for Twitter than for the published writing corpus, these words come from a more diverse set source languages and often include slang from other cultures (e.g., *ahre*; De Luca, 2021). In many ways, Twitter neologisms are less typical than the published writing ones, which could explain the differences in the neighbourhood statistics between the two domains.

The differences in the corpus construction also impact the experimental results: $\mathcal{D}_{\text{HISTORICAL}}^{\text{Pub}}$ contains 18 timesteps (decades from 1800s to 1980s), while $\mathcal{D}_{\text{HISTORICAL}}^{\text{Tw}}$ has only four (years from 2007 to 2010). As a result, the estimates of the frequency growth monotonicity measure (introduced originally for the published writing corpus; Ryskina et al., 2020b) are too noisy for Twitter (Figure 2, middle column). However, our frequency growth slope measure shows support for the demand hypotheses both on Twitter (Figure 2, top left) and in published writing (Figure 1, top left) for neighbourhoods in the static embedding space.

Finally, we should note that our definition of a neologism—a new form–meaning pair that spreads beyond its initial user(s) to a wider community—does not distinguish between true *broadening of use* and the *growth of the user community itself*. Especially for Twitter, the trends we see might be affected by how the platform’s user base changes. The increased use of South African slang words in English contexts could be due to an influx of users from South Africa, and the growing popularity of terms associated with K-pop (Korean popular music) could be explained by more of the platform’s users becoming interested in K-pop. Future work on social media neology could focus specifically on posts aimed at a more general audience (such as the *r/CasualConversation* subreddit), additionally evaluate the degree of a neologism’s social or geographic diffusion (Eisenstein et al., 2014; Grieve et al., 2018), or jointly consider the two domains explored in this paper: a word’s adoption by the published/mass media is a strong indicator of it entering the mainstream vocabulary.

Impact of embedding choice The experimental results for the demand hypothesis differ depending on whether static or contextual embeddings are used. The difference is especially notable for Twit-

ter, where the neighbours of the neologisms in the RoBERTa space seem to grow in frequency *less* than the neighbours of the control words (Figure 2, bottom right)—the opposite of what the hypothesis predicts. After inspecting the neighbourhoods of the neologisms, we find that the distance in the RoBERTa space is substantially affected by subword token overlap, especially if any of the subwords are rare. Words sharing a root tend to be closer in this space, which is useful for neologisms created by derivation or compounding, as is common in published writing (Table 3). However, Twitter abbreviations, blends, and creative spellings often contain unusual character sequences and end up being over-segmented by the RoBERTa tokenizer. For example, the nearest neighbours for *smol* are other spelling variants starting with *sm-* (*smthin*, *smtimes*, *smoooooth*) rather than semantically similar words like *cute*. This suggests that contextual embeddings may not be the best tool for studying social media neologisms specifically. In general, contextual embeddings might be better suited for a different operationalization of the hypotheses (e.g., using neighbourhoods of senses rather than words).

8 Conclusion

In this paper, we compare two semantic correlates of neology between two domains: historical published writing (books and articles) and social media (Twitter). We collect a large corpus of English tweets and extend our earlier methodology (Ryskina et al., 2020b) to this data in order to test two hypotheses about neology: that new form–meaning pairs emerge to fill semantic gaps (supply hypothesis) and that they emerge in topics that are rising in popularity (demand hypothesis). We operationalize the hypotheses using either static or contextual word embeddings, though we find the latter less suitable for social media data due to the impact of tokenization. For published writing, we reproduce our earlier results, finding evidence for both hypotheses. For Twitter, we find that though both hypotheses are supported, the evidence for the demand hypothesis is less conclusive. While it can be partially attributed to the differences in the corpus construction, we hypothesize that the creative tendencies of social media language use (evident in the neologism formation mechanisms prevalent on the platform) might be competing with the demand for word creation driven by the development of novel concepts in topics of growing popularity.

Limitations

As acknowledged in §7, there are limitations to extending our frequency-based framing of neology from books and newspapers to social media. While social media provides a unique opportunity to study linguistic innovation in a much wider and more diverse population of language users, the composition of this population can also change rapidly, and it can be difficult to tell whether the observed frequency effects are due to the neologism’s spread across communities of users or simply to the growth of the particular community it originated from. Public social media posts, despite being accessible to any user, are also not necessarily meant to be understood by a general audience, which further obscures the distinction between occasional in-group language play and more lasting and widespread language change.

We would also like to acknowledge several methodological caveats. Because of the strict neologism–control matching criteria, not all identified neologisms are matched to control words (§5.3) and therefore included in the final analysis, which may affect its findings. Certain differences in the design choices made for each of the two domains (the relative lengths of the HISTORICAL and the MODERN time periods, the choice to limit neologisms to nouns only or to allow other parts of speech) might also contribute to the observed differences. Finally, pretrained contextual embeddings, while commonly used in language change studies, inevitably depend on their original pretraining corpus, which is not specifically aligned with either of our domains or time periods; future work could explore using domain- and time-specific masked language models instead.

Acknowledgments

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute (www.vectorinstitute.ai/partnerships/current-partners/). We would like to thank Aria Haghghi for his comments during the conceptualization of the project and for the help with data collection, Eduard Hovy and Roger Levy for their feedback on an earlier version of this paper, Stas Kashepava for the contributions to the neologism formation analysis, and the two anonymous reviewers for their valuable suggestions.

References

- Jean Aitchison. 2001. *Language change: Progress or decay?* Cambridge University Press.
- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. [Contextual string embeddings for sequence labeling](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Aparna Ananthasubramaniam, David Jurgens, and Daniel M Romero. 2024. [Networks and identity drive the spatial diffusion of linguistic innovation in urban and rural areas](#). *npj Complexity*, 1(1):14.
- Aparna Ananthasubramaniam, Yufei ‘Louise’ Zhu, David Jurgens, and Daniel M. Romero. 2025. [Roles of network and identity in hashtag diffusion](#). In *Proceedings of the ACM on Web Conference 2025, WWW ’25*, page 3233–3246, New York, NY, USA. Association for Computing Machinery.
- Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. [A diachronic Italian corpus based on “L’Unità”](#). In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 369–374, Bologna, Italy. CEUR Workshop Proceedings.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O’Reilly Media Inc.
- Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. [Interpreting pretrained contextualized representations via reductions to static embeddings](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.
- Gerlof Bouma, Evie Coussé, Trude Dijkstra, and Nicole van der Sijs. 2020. [The EDGeS diachronic Bible corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5232–5239, Marseille, France. European Language Resources Association.
- Claire Bovern. 2019. [Semantic change and semantic stability: Variation is key](#). In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 48–55, Florence, Italy. Association for Computational Linguistics.
- Michel Bréal. 1904. *Essai de sémantique (science des significations)*. Hachette.
- Paul Cook and Suzanne Stevenson. 2010. [Automatically identifying the source words of lexical blends in English](#). *Computational Linguistics*, 36(1):129–149.
- Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013.

- No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 307–318, New York, NY, USA. Association for Computing Machinery.
- Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.
- Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.
- Natalia De Luca. 2021. El marcador conversacional “ahre” en el habla de estudiantes adolescentes de Buenos Aires: un estudio de sus usos en dos géneros conversacionales. *Revista Latinoamericana de Estudios del Discurso*, 21(2):27–48.
- Marco Del Tredici, Malvina Nissim, and Andrea Zaninello. 2016. Tracing metaphors in time through self-distance in vector spaces. In *3rd Italian Conference on Computational Linguistics, CLiC-it 2016 and 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2016*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.
- Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.
- Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13.
- David Francis, Ella Rabinovich, Farhan Samir, David Mortensen, and Suzanne Stevenson. 2021. Quantifying cognitive factors in lexical decline. *Transactions of the Association for Computational Linguistics*, 9:1529–1545.
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
- Jack Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics*, 46(4):293–319.
- Thomas Haider and Steffen Eger. 2019. Semantic change and emerging tropes in a large corpus of New High German poetry. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 216–222, Florence, Italy. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020. Predicting the growth of morphological families from social and linguistic factors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7273–7283, Online. Association for Computational Linguistics.
- John E. Hopcroft and Richard M. Karp. 1973. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on computing*, 2(4):225–231.
- Aniket Kali, Yang Xu, and Suzanne Stevenson. 2024. Cognitive factors in word sense decline. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.
- Andres Karjus, Richard A. Blythe, Simon Kirby, and Kenny Smith. 2020. Quantifying the dynamics of topical fluctuations in language. *Language Dynamics and Change*, 10(1):86–125.
- Andres Karjus, Richard A Blythe, Simon Kirby, Tianyu Wang, and Kenny Smith. 2021. Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, 45(9):e13035.
- Carmen Klaussner and Carl Vogel. 2018. A diachronic corpus for literary style analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Marie-Pauline Krielke, Luigi Talamo, Mahmoud Fawzi, and Jörg Knappen. 2022. Tracing syntactic change in the scientific genre: Two Universal Dependency-parsed diachronic corpora of scientific English and

- German**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4808–4816, Marseille, France. European Language Resources Association.
- Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziem, David Muchlinski, and Diyi Yang. 2024. **Silent signals, loud impact: LLMs for word-sense disambiguation of coded dog whistles**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12493–12509, Bangkok, Thailand. Association for Computational Linguistics.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. **Statistically significant detection of linguistic change**. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.
- Vivek Kulkarni, Yingtao Tian, Parth Dandiwal, and Steve Skiena. 2018. **Simple neologism based domain independent models to predict year of authorship**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 202–212, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Vivek Kulkarni and William Yang Wang. 2018. **Simple models for word formation in slang**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1424–1434, New Orleans, Louisiana. Association for Computational Linguistics.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. **Diachronic word embeddings and semantic shifts: a survey**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Natalie Lefkowitz and John S. Hedgcock. 2017. **Anti-language: Linguistic innovation, identity construction, and group affiliation among emerging speech communities**. In Nancy Bell, editor, *Multiple Perspectives on Language Play*, pages 347–376. De Gruyter Mouton.
- Alessandro Lenci and Magnus Sahlgren. 2023. *Distributional Semantics*. Cambridge University Press.
- Paul Lerner and François Yvon. 2025. **Towards the machine translation of scientific neologisms**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 947–963, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. **RoBERTa: A robustly optimized BERT pretraining approach**. *arXiv preprint arXiv:1907.11692*.
- Nikolay Malkin, Sameera Lanka, Pranav Goel, Sudha Rao, and Nebojsa Jojic. 2021. **GPT perdetry test: Generating new meanings for new words**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5542–5553, Online. Association for Computational Linguistics.
- Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. **Leveraging contextual embeddings for detecting diachronic semantic shift**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.
- Jane McCauley. 2006. **Technical combining forms in the third edition of the OED: Word formation in a historical dictionary**. In *Selected proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (HEL-LEX)*. Somerville, MA: Cascadia Proceedings Project, pages 95–104.
- Gretchen McCulloch. 2020. *Because Internet: Understanding the new rules of language*. Penguin.
- Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. **Quantitative analysis of culture using millions of digitized books**. *Science*, 331(6014):176–182.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013. **Distributed representations of words and phrases and their compositionality**. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.
- Filip Miletic and Sabine Schulte im Walde. 2025. **Modeling the evolution of English noun compounds with feature-rich diachronic compositionality prediction**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20071–20092, Vienna, Austria. Association for Computational Linguistics.
- James Milroy and Leslie Milroy. 2012. *Authority in Language: Investigating Standard English*. Routledge.
- Syrielle Montariol, Matej Martinc, and Lidia Pivovarov. 2021. **Scalable and interpretable semantic change detection**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.
- Martin Montgomery. 2008. *An introduction to language and society*. Routledge.
- David R. Mortensen, Valentina Izrailevitch, Yunze Xiao, Hinrich Schütze, and Leonie Weissweiler. 2024.

- Verbing weirds language (models): Evaluation of English zero-derivation in five LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17359–17364, Torino, Italia. ELRA and ICCL.
- C. Onelli, D. Proietti, C. Seidenari, and F. Tamburini. 2006. **The DiaCORIS project: a diachronic corpus of written Italian**. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).
- Kylie Pepler. 2013. Social media and creativity. In *The Routledge International Handbook of Children, Adolescents and Media*, pages 219–226. Routledge.
- Francesco Periti and Stefano Montanelli. 2024. **Lexical semantic change through large language models: a survey**. *ACM Comput. Surv.*, 56(11).
- Francesco Periti and Nina Tahmasebi. 2024. **A systematic comparison of contextualized word embeddings for lexical semantic change**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.
- Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020. **Will it unblend?** In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1525–1535, Online. Association for Computational Linguistics.
- Octavian Popescu and Carlo Strapparava. 2015. **SemEval 2015, task 7: Diachronic text evaluation**. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878, Denver, Colorado. Association for Computational Linguistics.
- Han Ren, Hai Wang, Yajie Zhao, and Yafeng Ren. 2023. **Time-aware language modeling for historical text dating**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13646–13656, Singapore. Association for Computational Linguistics.
- Everett M. Rogers and F. Floyd Schoemaker. 1971. *Communication of innovations: a cross-cultural approach*. The Free Press.
- Maja Rudolph and David Blei. 2018. **Dynamic embeddings for language evolution**. In *Proceedings of the 2018 World Wide Web Conference, WWW '18*, page 1003–1011, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Pablo Ruiz Fabo, Clara Martínez Cantón, Thierry Poibeau, and Elena González-Blanco. 2017. **Enjambement detection in a large diachronic corpus of Spanish sonnets**. In *Proceedings of the Joint SIGHUM*
- Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 27–32, Vancouver, Canada. Association for Computational Linguistics.
- Maria Ryskina, Matthew R. Gormley, and Taylor Berg-Kirkpatrick. 2020a. **Phonetic and visual priors for decipherment of informal romanization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8308–8319, Online. Association for Computational Linguistics.
- Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020b. **Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods**. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376, New York, New York. Association for Computational Linguistics.
- Kuleen Sasse, Carlos Alejandro Aguirre, Isabel Cachola, Sharon Levy, and Mark Dredze. 2025. **Making FETCH! happen: Finding emergent dog whistles through common habitats**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5709, Vienna, Austria. Association for Computational Linguistics.
- Ian Stewart and Jacob Eisenstein. 2018. **Making “fetch” happen: The influence of social and linguistic context on nonstandard word growth and decline**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370, Brussels, Belgium. Association for Computational Linguistics.
- Zhewei Sun and Yang Xu. 2022. **Tracing semantic variation in slang**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1313, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1).
- William Timkey and Marten van Schijndel. 2021. **All bark and no bite: Rogue dimensions in transformer language models obscure representational quality**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Eddie L. Ungless, Nina Markl, and Björn Ross. 2025. **Le\$bean or lesbian? A survey of marginalised users’ motivations for obfuscation on TikTok**. *Behaviour & Information Technology*, pages 1–26.
- Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle,

- Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. [Counting the bugs in ChatGPT’s wugs: A multilingual investigation into the morphological capabilities of a large language model](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.
- Winston Wu and David Yarowsky. 2020. [Computational etymology and word emergence](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.
- Quirin Würschinger, Mohammad Fazleh Elahi, Desislava Zhekova, and Hans-Jörg Schmid. 2016. [Using the web and social media as corpora for monitoring the spread of neologisms. the case of ‘rapefugee’, ‘rapeugee’, and ‘rapugee’](#). In *Proceedings of the 10th Web as Corpus Workshop*, pages 35–43, Berlin. Association for Computational Linguistics.
- Quirin Würschinger. 2021. [Social networks of lexical innovation. investigating the social dynamics of diffusion of neologisms on Twitter](#). *Frontiers in Artificial Intelligence*, 4.
- Aotao Xu, Charles Kemp, Lea Frermann, and Yang Xu. 2023. [Predicting strategy choice in word formation: A case study of reuse and compounding](#). In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.
- Aotao Xu, Christian Ramiro, and Yang Xu. 2019. [A predictability-distinctiveness trade-off in the historical emergence of word forms](#). In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 41.
- Yating Zhang, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. 2015. [Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 645–655, Beijing, China. Association for Computational Linguistics.
- Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, and Katsumi Tanaka. 2016. [The past is not a foreign country: Detecting semantically similar terms across time](#). *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807.
- Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. [NEO-BENCH: Evaluating robustness of large language models with neologisms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.
- ## A Differences from Ryskina et al.’s (2020b) experimental setup
- Our operationalization of the hypotheses (§5.4) builds on that of Ryskina et al. (2020b), with several differences:
- Since the differences in neighbourhood density between neologisms and controls can be dramatic, we measure it as $\log(|N_\tau(w)| + 1)$ (Eq. 1) rather than the raw number of neighbours $|N_\tau(w)|$.
 - We propose a more robust estimation for the frequency growth monotonicity measure: instead of estimating each neighbour word’s growth monotonicity and averaging, we measure the growth of the neighbourhood as a whole (Eq. 2) and estimate the monotonicity of that (Eq. 3).
 - To test the demand hypothesis more thoroughly, we also compute the regression slope of the frequency growth (Eq. 4) as an additional metric.
- We also use different criteria for neologism selection (§5.1) and neologism–control pairing (§5.3), use different hyperparameters for Word2Vec (§5.2), and additionally repeat all experiments with contextual embeddings (§5.2).
- ## B Data preprocessing and filtering
- ### B.1 Tweet preprocessing and tokenization
- At the preprocessing step, all tweets are lowercased and tokenized using NLTK’s TweetTokenizer (Bird et al., 2009). Using regular expression-based heuristics, we remove the following from all tokenized tweets: URLs; phone numbers; hashtags; tokens consisting exclusively of emoji, numbers, punctuation, or special characters; single-character tokens; and tokens longer than 50 characters. This leads to an extra 0.3% of all tweets being filtered out entirely. We use this tokenized data to:
- extract the vocabulary and estimate token frequencies, which are then used to select candidate neologisms (§5.1) and to pair neologisms with controls (§5.3);
 - train the static Word2Vec models (§5.2).
- Since RoBERTa uses its own tokenizer, we extract the contextual embeddings (§5.2) using the original, non-preprocessed tweet text.

However, some vocabulary words extracted by TweetTokenizer do not match RoBERTa’s token boundaries. For example, in

```
RT<USERNAME>:Parenting tip ...
```

TweetTokenizer considers :P to be a separate emoticon token, leading to arenting appearing in the historical vocabulary. Such tokenization mismatches are excluded from the historical vocabulary in all RoBERTa-based analyses.

B.2 Automatic Twitter neologism filtering

The initial list of neologism candidates obtained by thresholding the MODERN Twitter vocabulary by the first year of popular use (§5.1) consists of 3554 words. First, we refine the resulting list using part-of-speech tagging: for each word, we sample 100 tweets containing it at random and run them through the Flair English POS tagger⁷ (Akbik et al., 2018). Any potential neologisms for which the most frequent tag was NNP or NNPS (proper nouns), FW (foreign), CD (number) or NFP (superfluous punctuation) are discarded. Second, we filter out rare variants by removing any words that occur fewer than 500 times in the entire corpus. Finally, we remove any words whose frequency distribution is sparse and sharply peaked (usually associated with auto-generated, templatic tweets). For example, theweatherchannel has 130K occurrences in 2016 and none in 2017–2021, and all 2016 occurrences follow the same template:

```
Get Weather Updates from
theweatherchannel <TIMESTAMP>
```

The 938 words remaining after this step are additionally filtered manually, as described in the next section. The results with the non-filtered list of words are reported in §D.

B.3 Manual neologism filtering

Published writing We look up each of the 1000 potential neologisms extracted by Ryskina et al. (2020b) in the Oxford English Dictionary (OED).⁸ An OED entry typically provides etymological information as well as the usage statistics for each recorded distinct sense of the word. For each word, we note the year it was first used in its latest sense in any part of speech (e.g., *icon* in the computing sense) or as a part of its newest OED-recorded compound (such as *in-vitro fertilization* for *in-vitro*).

⁷<https://huggingface.co/flair/pos-english>

⁸<https://www.oed.com/>

We discard any words for which this year is earlier than 1900. For words not found in OED entries (e.g., *p-value*), we rely on either the first use date per the Merriam-Webster dictionary⁹ or the year of the word’s first appearance in the OED quotation bank. Finally, we discard the nine remaining words not found in either dictionary.

Twitter Starting with the list of 938 neologism candidates, obtained as described in §5.1, we look up each word’s context in our Twitter dataset, as well in the variety of online sources.

We design the following filtering rules:

- We discard foreign words for which no evidence of natural code-switched use was found in our corpus. These occur in tweets created by sharing from other websites (e.g., “... *via YouTube*”):

```
Need for Speed - Most Wanted
Soundtrack (Full) <URL> przez
@YouTube
```

We keep foreign words used in code-switched contexts:

```
Nha Bafethu the Messi saga is
difficult to accept
```

- We discard proper names, including names of people, works of art, products, and companies:

```
VOTE SUNGHOOON ON TIKTOK!
```

We keep derivations or abbreviations of product names that have gained common use (e.g., per Urban Dictionary):¹⁰

```
I just found a bunch of old
selfies/insta posts
```

We also keep product names converted into verbs:

```
using snapchat for the
memories >>> actually
snapchatting
```

- We discard typos:

```
DM US ON OUR INSTAGRAM
ACCOUNT
```

⁹<https://www.merriam-webster.com/>

¹⁰<https://www.urbandictionary.com/>

We use the Urban Dictionary to differentiate them from deliberate creative spellings, which we keep:

i literallt¹¹ cant

- We discard tokenization errors, such as lled being identified as a neologism candidate because of incorrectly splitting a token:

i forgot to censor k!lled

- We discard words which have been in use prior to 2000 per the Oxford English Dictionary (OED) and have not gained new senses since (e.g., unvaccinated). We rely on both the OED and the Urban Dictionary for identifying senses: for example, the OED lists only the currency-related sense of demonetization (in use since 1795), while the Urban Dictionary lists its newer content-related sense.¹²

C Neologism formation categorization

We primarily rely on the Oxford English Dictionary (OED) to categorize neologisms by how they were created. For non-dictionary words or ones where the etymological information is not provided, we reconstruct it using Wiktionary¹³ and other online resources.

The statistics for both corpora are reported in Table 3. We use the following categories:

- Abbreviation: clippings or shortenings (*information* → *info*), acronyms (*canola*), and initialisms (*aka*). Following the OED labels, we also include clipped compounds (*sitcom*) and loan-phrases shortened to single words (*caffè latte* → *latte*).
- Derivation: morphological derivatives, primarily words created by affixation (*light* → *ultralight*). Words that the OED annotates as created by back formation (*grungy* → *grunge*) are included in this category.
- Borrowing: loanwords from other languages, including cases where a borrowed root is combined with an English affix (Greek ‘*leptos*’ + *-in* → *leptin*). For Twitter neologisms, we include foreign words used in code-switched contexts in this category.

¹¹<https://www.urbandictionary.com/define.php?term=Literallt>

¹²<https://www.urbandictionary.com/define.php?term=Demonetization>

¹³<https://www.wiktionary.org/>

- Compound: words combined in their entirety (*bodysuit*, *business-to-business*) or words combined with affixes that the OED classifies as combining forms (such as *bio-*, *tele-*; McCauley, 2006).
- Blend: words combined in a way that alters at least one of their original forms (e.g., *cell(ular)* + *phone* → *cellphone*, *t(rans)-* + *am(ine)* + *oxy-* + *phen(ol)* → *tamoxifen*). Borrowed roots combined in such a way with English elements (Greek ‘*ergon*’ + (*econ*)omics) → *ergonomics*. The OED does not differentiate between blends and compounds, so we classify them manually.
- Spelling: orthographic alterations of an existing word or phrase (*gangster* → *gangsta*). Orthographic renderings of pronunciation or accent (*both of them* → *boffum*, *heart* → *hearteu*), onomatopoeia (*skrrt*), and keysmash (*sksksk*; McCulloch, 2020) are also included in this category.
- Sense: a commonly used word acquiring a new sense. While the OED provides both the original etymology (e.g., borrowing for *analog*) and the timelines for the various senses (e.g., 1941– for *analog* as ‘non-digital’), for many words (e.g., *browser*) it is hard to say whether the new sense was transferred onto an existing form or the form itself was re-coined from scratch. We chose to use this label (in addition to the original OED-annotated category) primarily for neologisms acquiring figurative senses.
- POS conversion: an existing word taking on a new part of speech (e.g., *aerobic*, adj. → *aerobics*, noun). If a word has multiple OED entries for different parts of speech listing different etymologies (e.g., derivation for *interface*, n. and conversion for *interface*, v.), we use labels from all the entries where the first use of the most recent sense is in 1900 or later. Where the OED combines parts of speech in the same entry, we follow the dictionary and do not use this label.

Many Twitter neologisms do not appear in the OED, so we annotated them manually, aiming for consistency with the OED categorization. The same word can fall into multiple categories (e.g., *bodycam* is both a compound and an abbreviation).

D Results for non-filtered neologisms

Figure 3 shows the results for the non-filtered list of published writing neologisms (755 neologism–control pairs). No suitable control words were found for the remaining 245 neologism candidates.

Figure 4 shows the results for the non-filtered list of Twitter neologisms (451 neologism–control pairs). No suitable control words were found for the remaining 487 neologism candidates.

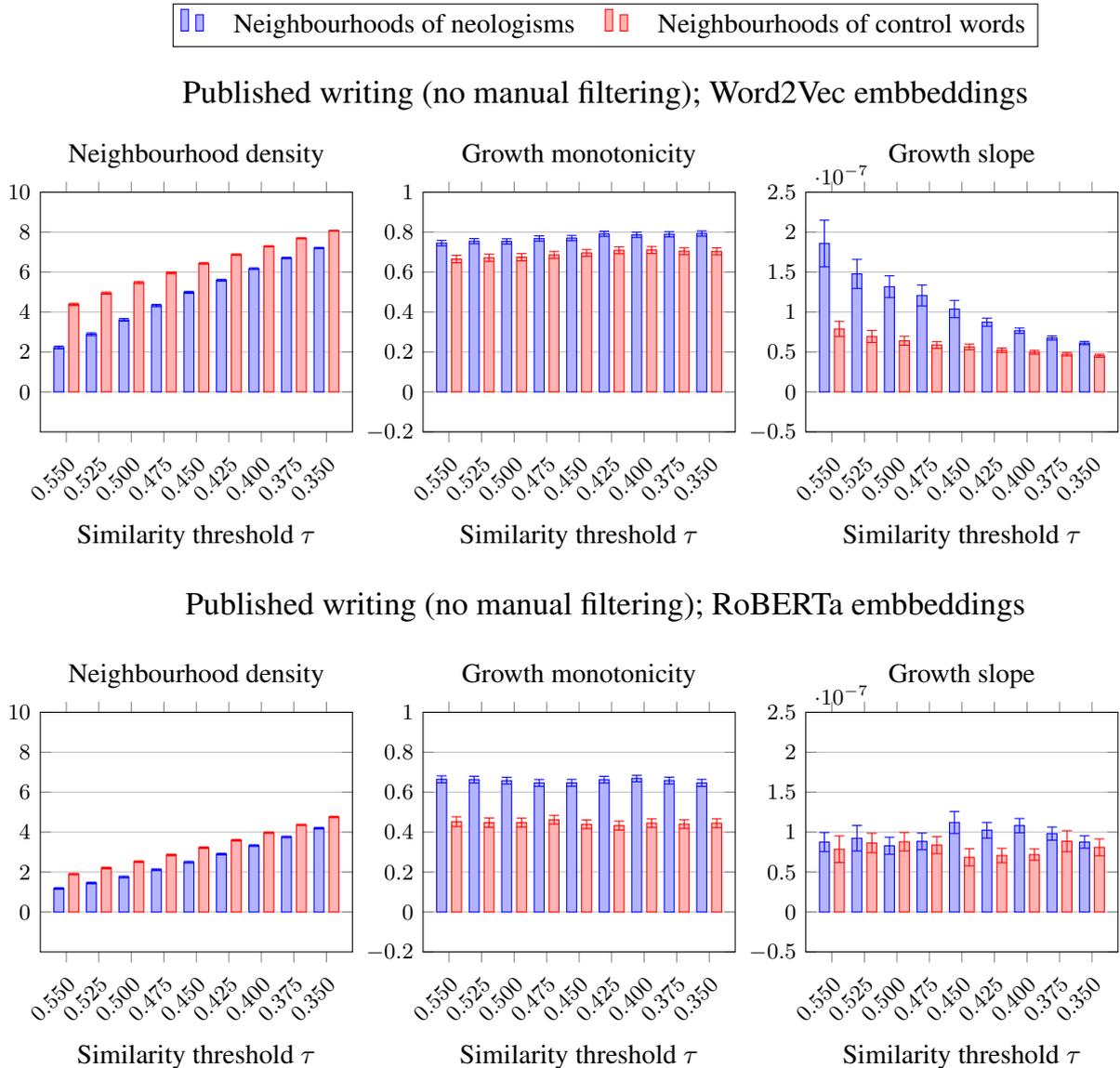


Figure 3: Experimental comparison between the neighbourhoods of neologisms (blue bars) and control words (red bars) in the **published writing** domain. Results are reported for 755 neologism–control pairs created from the original, **non-filtered neologism list** of 1000 candidate neologisms. The three plots in each row correspond to three measures: the number of HISTORICAL neighbours a word has (left), how monotonically these neighbours grow in frequency (centre), and the linear regression slope of their growth (right). The x-axis on all plots corresponds to the neighbourhood size (defined by the cosine similarity threshold τ). The top and bottom rows show the results with the static Word2Vec embeddings and the contextual RoBERTa embeddings respectively. Error bars represent standard error over words.

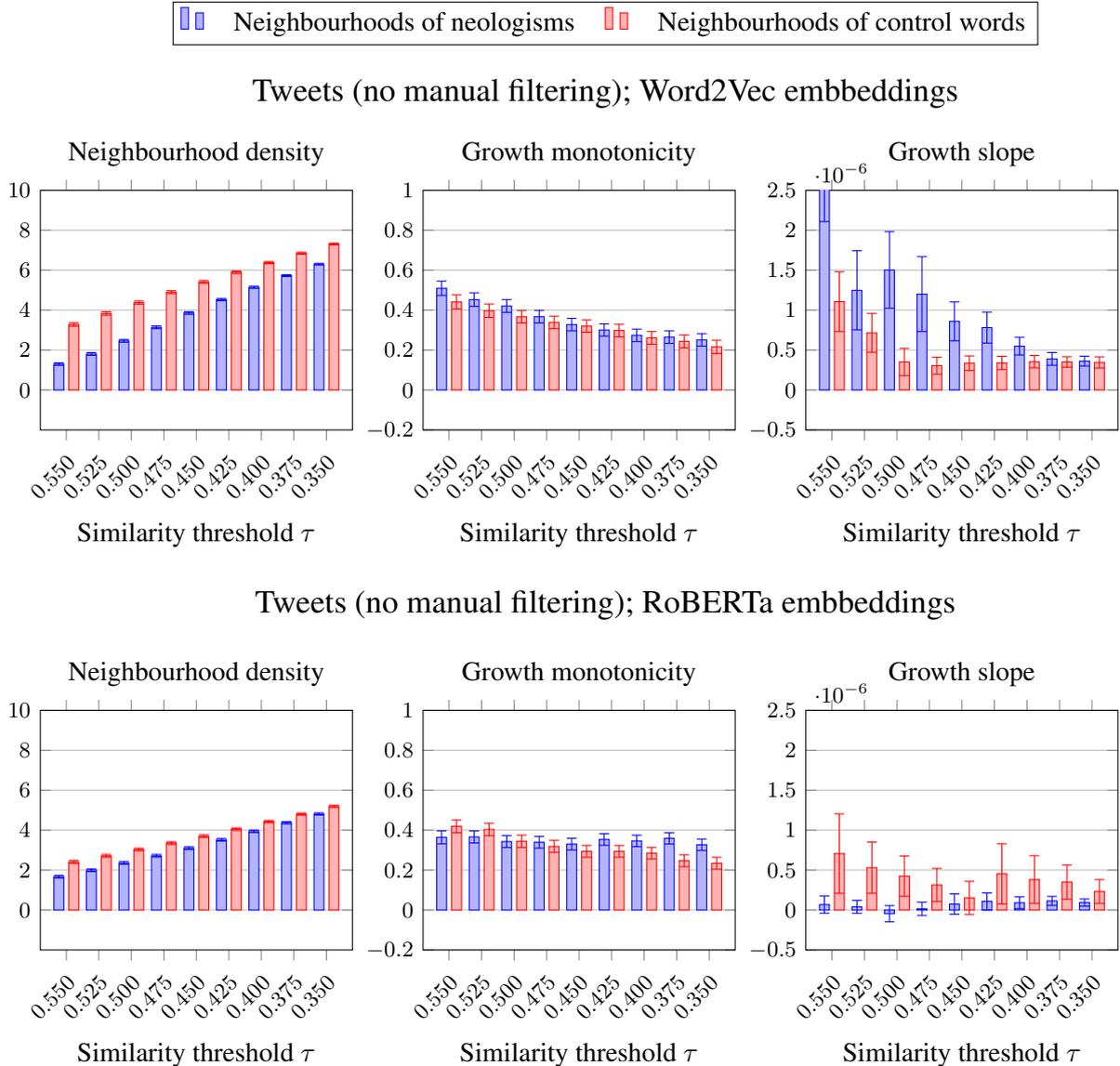


Figure 4: Experimental comparison between the neighbourhoods of neologisms (blue bars) and control words (red bars) in the **Twitter** domain. Results are reported for 451 neologism–control pairs created from the original, **non-filtered neologism list** of 938 candidate neologisms. The three plots in each row correspond to three measures: the number of HISTORICAL neighbours a word has (left), how monotonically these neighbours grow in frequency (centre), and the linear regression slope of their growth (right). The x-axis on all plots corresponds to the neighbourhood size (defined by the cosine similarity threshold τ). The top and bottom rows show the results with the static Word2Vec embeddings and the contextual RoBERTa embeddings respectively. Error bars represent standard error over words.

Author Index

- Aida, Taichi, 20
Aloraini, Afnan Mohammed, 1
- Baes, Naomi, 50
Batista-Navarro, Riza, 1
Berg-Kirkpatrick, Taylor, 162
Blum, Frederic, 75
- Da Silveira, Marcos, 110
De Pascale, Stefano, 97
de Sá, Jäder Martins Camboim, 110
Dinu, Anca Daniela, 35
Dinu, Liviu P, 35
Dubossarsky, Haim, 147
- Fedorova, Mariia, 87
- Geeraerts, Dirk, 97
Georgescu, Simona, 35
Gormley, Matthew R., 162
Goworek, Roksana, 147
- Haslam, Nick, 50
Heylen, Kris, 97
- Iacob, Alina, 131
- Komachi, Mamoru, 20
Kulkarni, Vivek, 162
Kutuzov, Andrey, 87
- Lee, Jooyoung, 110
Liang, Siyu, 27
List, Johann-Mattis, 75
- Mahowald, Kyle, 162
Mochihashi, Daichi, 20
Mortensen, David R., 162
- Nenadic, Goran, 1
- Ogiso, Toshinobu, 20
- Pruski, Cedric, 110
- Radchankava, Liudmila, 124
Ryskina, Maria, 162
- Schlegel, Viktor, 1
Speelman, Dirk, 97
- Takamura, Hiroya, 20
Tat, Bach Phan, 97
- Uban, Ana Sabina, 35, 131
Umarova, Khonzoda, 87
- Zerong, Zhaxi, 27