LREC 2026

NEOLOGY AND LARGE LANGUAGE MODELS

Workshop Proceedings

Editors
Valunaite Oleskeviciene Giedre and Giouli Voula

16 MAY 2026

ISBN 978-2-493814-70-8

EAN 9782493814708

Preface

Neology, the process of creating and disseminating new words, has a long tradition in lexicography, lexicology, corpus and computational linguistics, and sociolinguistics. Recently, with the advent and rapid diffusion of Large Language Models (LLMs), the analysis of lexical innovation has entered a new era. LLMs process linguistic data at an unprecedented scale, modeling language in ways that mirror, extend, and sometimes distort human lexical creativity (Qin et al., 2025; Hosseini-Kivanani, 2025). Undoubtedly, LLMs are powerful tools for detecting new words (Tomaszewska et al., 2025) and tracking their life cycle (Cartier, 2017) across domains, genres, and communities of use (Ichien et al., 2024), thus complementing and extending corpus-based approaches to neology. Moreover, LLMs not only absorb neologisms from their training data but also have the capacity to generate novel lexical items, metaphors, or hybrid forms in response to prompts (Iwamoto and Kanayama, 2024). This raises important questions about authority, legitimacy, and the mechanisms of linguistic innovation. From a sociolinguistic perspective, the interaction between LLMs and neology highlights both opportunities and risks. On one hand, LLMs could assist communities in documenting and systematizing emergent lexical forms, thus supporting language revitalization. On the other hand, their biases and uneven coverage may privilege neologisms in high-resource languages while under-representing or distorting innovation in smaller linguistic communities. Ultimately, this dynamic positions LLMs not just as tools for text processing but as active participants in shaping the future trajectories of language use and innovation. These developments open new perspectives for tracking linguistic creativity, modeling semantic shifts, and supporting lexicographic and terminological research.

The NeoLLM2026 workshop explores the intersection of LLMs and neology, focusing on—but not limited to—how cutting-edge computational methodologies leveraging LLMs can identify, generate, analyze, and evaluate new words and semantic shifts across languages and language varieties.

The invited speaker, Maciej Ogrodniczuk, from the Institute of Computer Science of the Polish Academy of Sciences, in his talk **From "Neologism of the Week" to "Neologism of the Year"**, addresses how Large Language Models (LLMs) can facilitate the study of lexical innovation by shifting the focus of neology towards continuous, data-driven observation and how LLMs can influence the detection, interpretation, and legitimisation of new words in contemporary discourse. At the empirical core of the study lies NeoN: an LLM-enhanced system that automatically detects, monitors, and analyzes neologisms in Polish, with a particular focus on its 'neologism of the week' module. Combining frequency dynamics, contextual evidence, and automated definition and categorization, the system documents which new words appear in discourse and how they gain or lose relevance over time. Situating NeoN within a wider cross-disciplinary framework, the talk reflects on how LLM-based approaches can complement traditional lexicography and support linguistic documentation, offering new methodological pathways for studying linguistic change. The talk will also explore whether the 'neologism of the year' is merely a linguistic artifact or a socio-technical construct, exposing the evolving interplay between human creativity, media dynamics and artificial intelligence.

Organizing Committee

Florentina Armaselu (University of Luxembourg)

Voula Giouli (Aristotle University of Thessaloniki)

Barbara Lewandowska-Tomaszczyk (University of Applied Sciences in Konin)

Chaya Liebeskind (Jerusalem College of Technology)

Barbara McGillivray (King's College London)

Giedre Valunaite Oleskeviciene (Mykolas Romeris University)

Scientific Committee

Florentina Armaselu (University of Luxembourg, Luxembourg)

Giorgio Maria Di Nunzio (University of Padua, Italy)

Radovan Garabík (Ludovit Stur Institute of Linguistics, Slovak Republic)

Voula Giouli (Aristotle University of Thessaloniki)

Anas Fahad Khan (CNR-Istituto di Linguistica Computazionale "Antonio Zampolli," Italy)

Barbara Lewandowska-Tomaszczyk (University of Applied Sciences in Konin)

Chaya Liebeskind (Jerusalem College of Technology, Israel)

Elpida Loupaki (Aristotle University of Thessaloniki, Greece)

Barbara McGillivray (King's College London)

Liudmila Mockiene (Mykolas Romeris University, Lithuania)

Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences, Poland)

Atul Kr. Ojha (University of Galway, Ireland)

Ana Ostroški Anić (Institute of Croatian Language and Linguistics, Croatia)

Marko Robnik-Šikonja (University of Ljubljana, Slovenia)

Purificação Silvano (University of Porto, Portugal)

Enriketa Sogutlu (Beder University, Albania)

Ranka Stankovic (University of Belgrade, Serbia)

Giovanni Luca Tallarico (University of Verona, Italy)

Giedre Valunaite Oleskeviciene (Mykolas Romeris University, Lithuania)

Andrius Utka (Vytautas Magnus University)

Federica Vezzani (University of Padua, Italy)

Table of Contents

Workshop Program

Saturday, 5/16/2026

14:00–18:00 **Session Workshop: Neology and Large Language Models**
Room: Room 9
Chairs: Giedre Valunaite Oleskeviciene, Voula Giouli

14:40–15:00 *From 124 Million Tokens to 1,021 Neologisms: A Large-Scale Pipeline for Automatic Neologism Detection*
Diego Rossini and Lonneke van der Plas

15:00–15:20 *High Resource Bias in AI-Driven Neology: Structural Inequality in Lexical Innovation*
Wajdi Zaghouani

15:20–15:40 *Do LLMs Know What Luxembourgish Borrows? Probing Lexical Neology in Low-Resource Multilingual Models*
Nina Hosseini-Kivanani

15:40–16:00 *Lexical Innovation in Business Colour Idioms: Evidence from Large Language Models in Five Languages*
Giedre Valunaite Oleskeviciene, Ágnes Abuczki, Ganit Richter, Berat Ujkani, Vera Moitinho de Almeida and Pedro Madeira

16:30–16:50 *Where in Semantic Space Do Spanish Neologisms Emerge?*
Bianca Delgado and Shira Wein

16:50–17:10 *Assessing the Pragmatic Competence of LLMs Regarding Novel Discourse Markers in Digital Communication*
Ágnes Abuczki and Giedre Valunaite Oleskeviciene

17:10–17:30 *A Comparative Evaluation of Semantic Ambiguity Detection in Two LLMs*
Lili Tamas

17:30–17:50 *LLM-Based Frame and Stance Annotation for 19th-Century Rumour Discourse in US and UK Newspapers*
Wanshu Zhang

From 124 Million Tokens to 1,021 Neologisms: A Large-Scale Pipeline for Automatic Neologism Detection

Diego Rossini, Lonneke van der Plas
Università della Svizzera italiana (USI)
Lugano, Switzerland
{diego.rossini, lonneke.vanderplas}@usi.ch

Abstract

We present a scalable, modular pipeline for automatic neologism detection that combines rule-based filtering with LLM classification. The pipeline is grounded in two complementary word-formation frameworks, grammatical and extra-grammatical morphology, which jointly define the scope of what counts as a neologism and inform a four-class classification scheme (NEOLOGISM, ENTITY, FOREIGN, NONE). While designed to be modular and transferable at the architectural level, the pipeline is instantiated on 527 million English-language Reddit posts spanning 2005–2024. From this corpus, we extract 124.6 million unique tokens and reduce them by over 99.99% to yield 1,021 neologism candidates, a set small enough for manual expert verification. Multiple LLMs independently classify each candidate via majority vote, with a final verification step, revealing substantial cross-model disagreement and highlighting the challenge of operationalizing neologism detection at scale. Manual annotation of all 1,021 candidates confirms that 599 (58.7%) are genuine lexical innovations. The pipeline code, vocabulary compilation scripts, and the annotated candidate list are available at https://github.com/DiegoRossini/neologism-pipeline.

Keywords: neologism detection, lexical innovation, Reddit, large language models, rule-based filtering, computational neology

1. Introduction

Although the study of neologisms has deep roots in linguistics (Guilbert, 1975; Rey, 1976), their automatic detection is a comparatively recent task. Computational approaches only became feasible once large machine-readable corpora were available in the 1990s (Renouf, 1993; Cabré and de Yzaguirre, 1995). Since then, a number of web-based platforms have been developed for neologism identification (Kerremans et al., 2012; Cartier, 2017; Klosa-Kückelhaus and Lüngen, 2018). These systems typically depend on static exclusion dictionaries and language-specific resources, and require manual expert verification of their output (Brasolin et al., 2023; Cartier, 2017; Tomaszewska et al., 2025). The key challenge for any detection pipeline is therefore to reduce the candidate set to a size where such verification is feasible.

More recently, social media data has attracted increasing attention as a source for studying lexical innovation, given the volume, diversity, and informality of user-generated content (Grieve et al., 2018; Würschinger, 2021). However, the same characteristics that make these platforms attractive also pose a challenge for neologism detection: in a large corpus, the vast majority of tokens absent from standard dictionaries are not neologisms but typos, misspellings, concatenated strings, code fragments, or foreign-language material. In the dataset used in this study, 527 million English-language Reddit posts spanning 2005–2024 (Baumgartner et al., 2020; Watchful1, 2025), we extract 124.6 million unique tokens, which the pipeline reduces by over 99.99% to yield 1,021 neologism candidates. Classifying each of the 124.6 million unique tokens individually with an LLM would be computationally infeasible, which motivates a multi-stage filtering approach that progressively narrows the set before classification.

In this paper, we present a pipeline for large-scale neologism detection that combines deterministic rule-based filtering with LLM-based classification. The rule-based stages progressively reduce the candidate set; the LLM stage then classifies surviving candidates into four categories: ENTITY, NEOLOGISM, FOREIGN, or NONE. Multiple LLMs independently classify each candidate, and only those receiving a majority vote are retained. A final verification step can then confirm or discard the output of the preceding models. Our contributions are: (1) a scalable, modular pipeline for neologism detection from social media corpora, grounded in word-formation theory (§3), whose output illustrates a range of grammatical and extra-grammatical word-formation processes (§7.1); (2) a comparative evaluation of multiple LLMs on a four-class neologism classification task; and (3) a detailed manual analysis of all 1,021 pipeline output candidates, including gold annotation, error analysis by category (§7.2), and classification of detected neologisms along word-formation processes (§7.1).[1]

[1]The pipeline code, vocabulary compilation scripts, and the annotated candidate list are available at https://github.com/DiegoRossini/neologism-pipeline.

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 1–15
16 MAY 2026.

2. Related Work

The dominant paradigm for automatic neologism detection remains the *exclusion dictionary method*: a token is flagged as a candidate neologism if it does not appear in one or more reference lexicons (Renouf, 1993; Cabré and de Yzaguirre, 1995). This principle underpins the major detection platforms developed over the past two decades, including the NeoCrawler (Kerremans et al., 2012, 2018), which monitored English-language websites for previously unattested forms, Néoveille (Cartier, 2017), which adopted a similar architecture for multiple languages, and the IDS Neologismenwörterbuch (Klosa-Kückelhaus and Lüngen, 2018), a continuously updated German neologism dictionary backed by corpus monitoring. The NeoCrawler was formally decommissioned in 2020 (Q. Würschinger, personal communication, 2025), illustrating the fragility of long-term tool availability. All three systems depend on language-specific resources that limit portability across languages and corpora. Our pipeline adopts the same foundational exclusion principle but separates language-specific resources from the architectural design: the sequence of filtering stages is pre-determined, while the resources they operate on (reference vocabularies, phonotactic rules, frequency dictionaries) should be substituted or adapted for each target language.

Beyond pure exclusion lookup, Falk et al. (2014) trained an SVM on French newspaper candidates using form-related, morpho-lexical, and thematic features, demonstrating the value of semantic context for neologism classification. Our pipeline follows a similar two-stage logic, but delegates the classification step to prompted LLMs rather than to feature-based classifiers.

As corpora drawn from social media have grown in size, so has the difficulty of the candidate extraction step itself. Grieve et al. (2018) identified 54 emerging words from 8.9 billion tokens of geolocated American Twitter data by tracking frequency increases over time and filtering manually. Mahler (2020) applied a comparable frequency-based methodology to Reddit, and Würschinger (2021) demonstrated how network metrics capture properties of lexical innovation that frequency measures alone cannot reveal. Brasolin et al. (2023) and Spina et al. (2024) extracted candidates from millions of geolocated Italian tweets through exclusion filtering and manual distillation, identifying hundreds of unattested word forms. A recurring challenge across these studies is that the vast majority of tokens absent from standard dictionaries are not neologisms but typos, misspellings, code fragments, or foreign-language material.

The most directly comparable recent system is NeoN (Tomaszewska et al., 2025), a multi-layered pipeline for Polish that combines frequency analysis, structural constraints, reference corpus checking, and spelling error detection with an LLM-based final filter, demonstrating that LLMs can serve as effective precision boosters after rule-based prefiltering. Our work differs from NeoN in several respects: we ground the pipeline in two complementary word-formation frameworks (§3) that define the scope of each category; we adopt a four-class taxonomy rather than a binary filter; we use a multi-model majority-vote scheme with independent verification rather than a single LLM; and we provide a detailed manual analysis of all pipeline output, including error categorisation and classification by word-formation process.

3. Theoretical Foundations

Any neologism detection pipeline presupposes an operational definition of what counts as a neologism. This section presents the two word-formation frameworks that jointly inform the design of our classification scheme and, in particular, determine the scope of the NEOLOGISM label assigned by the LLM stage (§4).

3.1. Grammatical Word Formation

Štekauer's onomasiological theory (Štekauer, 1998; Štekauer, 2001; Štekauer, 2005) models word formation as a top-down, need-driven process: a speaker identifies a concept lacking a conventional expression and coins a new naming unit by selecting an onomasiological type — a structural pattern that maps conceptual content onto morphological form. The process is *grammatical* in the sense that, given a naming need, the resulting formation is constrained by the productive onomasiological types available in the language.

A central consequence of this framework concerns nonce-formations. Against the view that these are deviant, context-dependent, and inherently non-lexicalisable coinages (cf. Hohenhaus, 1998), Štekauer (2002) argues that nonce-formations are regular products of the Word-Formation Component, generated by the same productive rules as any other naming unit. What distinguishes them is not structural deviance but *lifecycle stage*: a nonce-formation is a neologism at the earliest point between coinage and dissemination, and whether it subsequently becomes institutionalised or falls out of use is an empirical matter that cannot be predicted at the time of coining. The notion of "nonce-formation" as a structurally distinct category thus collapses into a temporal label.

For the pipeline, this entails that tokens attested infrequently in the corpus cannot be excluded as

non-neologisms on formal grounds alone, since nonce-formations are structurally indistinguishable from formations that will eventually become established. The frequency threshold (§4.5) is accordingly designed to filter noise rather than to impose a lexicalisation requirement. However, Štekauer's model is explicitly limited to rule-governed formation: processes such as blending, clipping, and acronymy, whose output cannot be derived from productive onomasiological types, fall outside the Word-Formation Component and are relegated to the Lexicon (Štekauer, 2001). The following framework addresses precisely this gap.

3.2. Extra-Grammatical Word Formation

Mattiello (2013) addresses those processes that fall outside the scope of grammatical morphology: clippings, blends, acronyms, abbreviations, and other formations whose input does not allow prediction of a regular output through any rule-based model of word formation. Within the framework of Natural Morphology (Dressler, 2000), these are classified as *extra-grammatical*, distinct from both core grammatical morphology (rule-governed, productive) and marginal morphology (partially regular). Although traditionally marginalised for their irregularity and unpredictability, Mattiello (2013) demonstrates that extra-grammatical processes are productive in their own right, particularly in informal registers, and that they comply with criteria of well-formedness and contextual suitability. The driving mechanism is *analogy* rather than rule application (Mattiello, 2013, 2017; Arndt-Lappe, 2015): new formations arise by modelling on existing words, either individually (surface analogy) or through recurrent patterns (analogy via schema). Mattiello (2017), following Booij (2010) and Plag (1999), extends this mechanism to playful coinages such as *doggo* (← *dog*), previously excluded as *expressive morphology*, i.e. playful or affective modifications of existing words (Zwicky and Pullum, 1987). Moreover, the boundary between extra-grammatical and grammatical morphology is not fixed, as formations that originate as creative coinages can over time become regular and productive (Körtvélyessy et al., 2022, 2021).

For the pipeline, this entails that the NEOLOGISM class must be broad enough to encompass both grammatical and extra-grammatical formations. Taken together, the two frameworks define the theoretical scope of the positive labels used in this study: the four-class classification scheme presented in §4 operationalises this joint definition, with the NEOLOGISM and ENTITY classes capturing genuine lexical innovations and FOREIGN and NONE isolating non-neologistic material that rule-based filtering alone cannot eliminate.

4. Methodology

The pipeline is designed to be modular: the sequence of filtering stages is pre-determined, but the resources each stage operates on (reference vocabularies, phonotactic rules, frequency dictionaries) are language-specific and must be substituted for each target language. The instantiation described below targets English.

4.1. Tokenization

Raw texts are tokenized using a spaCy language model, with named entity recognition, dependency parsing, and lemmatization disabled for efficiency. The choice of model depends on the target language. A corpus-specific preprocessing step removes or replaces non-lexical content before tokenization: for social media corpora, this may include URLs, platform-specific references, user mentions, hashtags, emojis, and non-ASCII characters; for other corpus types, different noise patterns (e.g., markup tags, metadata fields) may require analogous treatment. Punctuation, stopwords, and whitespace tokens are discarded, and all remaining tokens are lowercased.

4.2. Vocabulary Filtering

Tokens present in a reference vocabulary compiled exclusively from sources predating a chosen cutoff date are filtered out on the assumption that they represent established lexical items rather than neologisms. The cutoff defines an observation window over which the lifecycle of detected neologisms can be tracked. The pipeline accepts one or more reference vocabularies; when multiple sources are available, combining them reduces the risk of false positives caused by gaps in any individual lexicon. The composition can be tailored to the target language and corpus: for social media data, it may include platform-specific vocabulary, crowdsourced slang dictionaries, and encyclopaedic entries alongside standard lexicons, while for more formal corpora, curated dictionaries and domain-specific terminologies may suffice. Any token found in the reference vocabulary is excluded from further processing.

4.3. Pattern-based Cleaning

Tokens surviving the vocabulary filter are subjected to pattern-based rules designed to remove noise that no dictionary would capture. These rules fall into two categories. The first is language-independent: tokens must be purely alphabetic and fall within a configurable length range, while those exhibiting excessive character repetition, low character entropy, or repeated character sequences are

discarded as likely keyboard spam or encoding artefacts. The second category is language-specific and must be adapted to the target language: this includes phonotactic constraints (e.g., implausible consonant or vowel clusters), expressive spelling variants (e.g., elongated interjections, laughter patterns), and corpus-specific noise patterns (e.g., placeholder or template text).

4.4. Typo and Concatenation Detection

Corpora, particularly those drawn from social media, frequently contain misspellings and tokens formed by words concatenated without spaces. Neither constitute lexical innovations, yet both survive vocabulary filtering because they do not match any reference entry. The pipeline applies SymSpell (Garbe, 2012), a symmetric delete spelling correction algorithm with support for multiple languages, to detect both cases against a reference frequency dictionary. A token is flagged as a typo if it falls within a configurable edit distance of a high-frequency entry in the dictionary, and as a concatenation if it can be segmented into two or more parts each appearing in the same dictionary. Minimum character length thresholds prevent spurious matches on short tokens, and a conservative maximum edit distance ensures that only tokens closely resembling high-frequency dictionary entries are flagged, so that morphologically complex forms such as compounds or blends are unlikely to be flagged; those that are can be recovered by the frequency-based reintegration mechanism (§4.5).

4.5. Frequency Threshold and Reintegration

Tokens previously excluded as typos or concatenations (§4.4) are reconsidered if they meet a configurable frequency threshold. If a token resembles a misspelling or a segmentable string yet recurs frequently in the corpus, it is unlikely to be an error, and its reintegration prevents genuine coinages from being prematurely discarded.

Candidates occurring fewer than the frequency threshold are excluded. While this introduces a tension with the theoretical position outlined in §3.1, where nonce-formations are treated as legitimate neologisms regardless of their frequency, the constraint is computational rather than theoretical: when too many candidates survive the rule-based filters, manual or LLM-based verification becomes infeasible. The threshold can be adjusted or omitted entirely depending on corpus size and available resources. In practice, nonce-formations attested below the threshold are lost; however, the threshold is not designed to impose a lexicalisation requirement but to separate deliberate, repeated use from accidental variation, since typos and random strings rarely recur at scale.

4.6. Foreign Language Detection

Depending on the target language, the corpus may contain substantial material from other languages that survives vocabulary filtering. The pipeline applies the Lingua language detector (Stahl, 2022) to flag and filter tokens identified as belonging to a language other than the target. A configurable confidence threshold controls how aggressively tokens are filtered. Tokens whose confidence score falls below the threshold, for instance due to mixed Tagalog–English morphology or orthographic overlap with the target language, are retained and delegated to the LLM classification stage, which includes a dedicated FOREIGN category. The stage does not distinguish foreign-language noise from loanwords entering the target language; this limitation is discussed in the Limitations section.

4.7. LLM Classification

Candidates surviving the filtering stages are classified into four categories using large language models. The taxonomy reflects the theoretical scope established in §3: NEOLOGISM (new words, slang, or words derived from proper nouns, encompassing both grammatical and extra-grammatical formations); ENTITY (proper nouns such as people, companies, brands, products, or places); FOREIGN (words from other languages that escaped the language detection stage, §4.6); and NONE (residual noise including usernames, typos, programming terms, and unclear cases).

Both NEOLOGISM and ENTITY constitute lexical innovations: in Štekauer's onomasiological framework, word-formation is a naming response to newly salient extra-linguistic referents, and proper nouns denoting emerging social or cultural entities qualify as newly established naming units. The two classes are kept separate for analytical purposes, facilitating comparison with standard NER categories in downstream applications. Classification is performed in two stages. First, multiple LLMs independently classify each candidate token. Labels are aggregated via majority vote: a token receives a label only if the majority of models agree; otherwise it is marked UNKNOWN. Second, an additional model verifies each label and produces the final output. The choice and number of models is configurable; using multiple architectures trained on different data reduces idiosyncratic misclassifications, while the independent verification step provides an additional quality control layer at minimal additional cost.

5. Experimental Setup

This section describes the instantiation of the pipeline for English-language neologism detection on Reddit data. All language-specific resources, parameters, and model choices reported below can be substituted for other languages or corpora.

5.1. Corpus

The corpus consists of Reddit submissions and comments spanning January 2005 to December 2024, drawn from the Pushshift archive (Baumgartner et al., 2020; Watchful1, 2025). After excluding deleted posts, removed content, and non-textual submissions, the dataset comprises approximately 527 million posts. Although the corpus is predominantly English, it contains multilingual content, most notably Taglish (Tagalog–English code-switching) in Filipino-oriented subreddits, as well as posts in Portuguese, Spanish, French, German, and other languages.

5.2. Tokenization

We use spaCy's `en_core_web_lg` model with named entity recognition, dependency parsing, and lemmatization disabled for efficiency. URLs, subreddit references (`r/\w+`), user mentions (`u/\w+`), and hashtags are replaced with placeholder tokens; emojis and non-ASCII characters are removed. Punctuation, stopwords, and whitespace tokens are discarded, and all remaining tokens are lowercased. The tokenization stage yields 124.6 million unique token types.

5.3. Reference Vocabularies

The reference vocabulary is compiled exclusively from pre-2015 sources, establishing a ten-year observation window (2015–2024) over which newly emerged tokens can be identified. The combined vocabulary comprises 16.3 million unique surface forms drawn from six sources (Table 1). Using multiple independently compiled resources reduces the risk of false positives caused by gaps in any individual lexicon: Reddit and Urban Dictionary (Urban Dictionary, 2025) cover informal register, Wikipedia titles (Wikimedia Foundation, 2015a) capture named entities and technical terminology, while WordNet (Princeton University, 2011), Wiktionary (Wikimedia Foundation, 2015b), and NoSlang (5.5K tokens, obtained with permission from the site owner) provide baseline lexical coverage.

5.4. Filtering Parameters

Pattern cleaning. Tokens must be purely alphabetic and between 3 and 20 characters in length. The English-specific rules filter tokens starting with double vowels (*aa*, *ee*, *ii*, *oo*, *uu*), implausible consonant clusters, expressive variants (*hahaha*, *yeaah*, *ughh*), repeated character sequences, and Lorem Ipsum placeholder words. Tokens exceeding six characters with two or fewer unique characters are discarded as keyboard spam. The full rule set is available in the project repository.

Typo and concatenation detection. SymSpell (Garbe, 2012) is configured with a maximum edit distance of 2 and a frequency dictionary compiled from Reddit pre-2015 token frequencies and WordNet. A token is flagged as a typo if its closest match in the dictionary has edit distance 1–2 and frequency above 100; minimum token length for typo checking is 5 characters. Concatenation detection applies word segmentation on tokens of at least 6 characters, flagging those that segment into two or more parts all present in the frequency dictionary. Genuine compounds flagged at this stage can be recovered by the reintegration mechanism described below.

Frequency threshold. The minimum occurrence threshold is set to 100 (§4.5). Tokens previously flagged as typos or concatenations are reintegrated if they meet this threshold.

Foreign language detection. The Lingua language detector (Stahl, 2022) is applied with a confidence threshold of 0.75 across 47 languages, removing 33,959 tokens (16.3% of the 208,932 candidates at that stage).

All thresholds reported above were set based on preliminary experimentation; a discussion of their limitations is provided in the Limitations section.

5.5. LLM Classification

The three open-source models—Qwen 2.5 72B, LLaMA 3.3 70B, and Mistral Large 2 123B—independently classify each candidate; labels are aggregated via majority vote (§4.7). Claude 4.5 Haiku serves as an independent verification source and does not participate in the vote.

All open-source models use few-shot prompting with eight labelled examples spanning the four classes and up to three contextual sentences per candidate, drawn from diverse subreddits. Claude 4.5 Haiku classifies tokens with the same contextual examples as the other models. The full prompt templates are provided in Appendix A.

5.6. Computational Setup

All experiments were run on a single multi-GPU server with 500 GB RAM and 4 GPUs (120 GB

Reference Vocabularies (all pre-2015)		
Source	**Tokens**	**Coverage**
Reddit pre-2015	10.5M	Informal, platform jargon
Wikipedia titles	4.4M	Entities, technical terms
Urban Dictionary	1.5M	Slang
Wiktionary	554K	Morphological variants
WordNet 3.1	147K	Core vocabulary
NoSlang	5.5K	Chat abbreviations
Total	**16.3M**	

Table 1: Reference vocabularies and primary coverage.

Stage	**Remaining**
Tokenization	124,593,754
Vocabulary lookup	113,909,871
Pattern cleaning	23,955,763
Concatenation detection	10,793,055
Typo detection	7,065,796
Freq. threshold + reintegration	208,932
Foreign language detection	174,973
Majority vote (NEOLOGISM)	10,499
Haiku verification	1,021

Table 2: Filtering cascade: candidates remaining after each stage.

each). The open-source models were sharded across all four GPUs in bfloat16 precision. Claude 4.5 Haiku was accessed via the Anthropic Batch API.

On the described hardware, the ideal critical path is approximately 50–65 hours (∼2–3 days): tokenization of 527 million posts accounts for 18–24 hours, vocabulary filtering and context retrieval for ∼9 hours, and sequential LLM inference over three models for 22–30 hours (40–50% of total compute). Running the three models in parallel on separate nodes would reduce the total to ∼38–49 hours.

6. Results

6.1. Filtering Cascade

Table 2 reports the number of candidate tokens surviving each pipeline stage. The rule-based stages reduce the initial 124.6 million unique tokens by 99.86%, yielding 174,973 candidates for LLM classification. The most aggressive single stage is pattern cleaning, which removes 90 million tokens (72.2% of the input at that point), followed by concatenation detection (13.2 million concatenated tokens) and vocabulary lookup (10.7 million known words). The frequency threshold eliminates a further 6.9 million low-frequency tokens. Of the tokens previously excluded as typos or concatenations, 118,544 meet the frequency threshold and are reintegrated into the candidate pool (§4.5).

6.2. LLM Classification and Inter-Model Agreement

The three open-source models independently classified all 174,973 tokens. Table 3 reports their label distributions, revealing substantial cross-model disagreement. LLaMA is the most aggressive NEOLOGISM predictor (12.2%, nearly double the other two models), while Mistral is the most conservative overall, assigning NONE to 59.9% of tokens. Qwen detects the most foreign-language material (22.2%).

Unanimous agreement across all three models is reached for only 45.8% of tokens (80,220); 48.4% are decided by a 2-out-of-3 majority, and 5.8% (10,134) result in three-way ties, conservatively resolved to NONE. These complementary biases validate the ensemble design: no single model would achieve the same coverage.

The majority vote produces 10,499 NEOLOGISM candidates (6.0%), 47,276 ENTITY (27.0%), 33,159 FOREIGN (19.0%), and 84,039 NONE (48.0%).

6.3. Haiku Verification

Claude 4.5 Haiku independently classified the same tokens with the same contextual examples. Applied as a verification filter to the 10,499 majority-vote NEOLOGISM candidates, Haiku confirmed 897 as NEOLOGISM (8.5%), relabeled 124 as ENTITY (1.2%), and rejected 9,478 to NONE (90.3%). This high rejection rate is driven primarily by model conservatism rather than prompt design or category confusion. Haiku receives the same multi-context prompts as the open-source models, yet assigns ENTITY to only 124 of 174,973 tokens (0.07%), compared to 47,276 from the majority vote, effectively defaulting all uncertain cases to NONE as instructed by the prompt. Across all tokens, it assigns NONE to 89.4%. This pattern places Haiku at the extreme end of a conservatism spectrum already visible among the open-source models, where Mistral (59.9% NONE) is markedly more conservative than Qwen (37.9%) and LLaMA (37.7%). Table 3 reports the full label distribution across all four models and the majority vote. The most striking pattern is Haiku's near-total rejection of the ENTITY class: of 47,276 majority-vote entities, none are confirmed and 97.3% are relabeled NONE. The verification stage thus acts as a strict precision filter, reducing the candidate set from 10,499 to 1,021.

6.4. Gold Standard Evaluation

The first author manually annotated all 1,021 pipeline output candidates using the same four-

Label	Qwen 72B	Mistral 123B	LLaMA 70B	Maj. vote	Haiku
NEOLOGISM	13,661 (7.8%)	11,493 (6.6%)	21,353 (12.2%)	10,499 (6.0%)	897 (0.5%)
ENTITY	56,144 (32.1%)	32,311 (18.5%)	55,088 (31.5%)	47,276 (27.0%)	124 (0.1%)
FOREIGN	38,851 (22.2%)	26,441 (15.1%)	32,625 (18.6%)	33,159 (19.0%)	17,506 (10.0%)
NONE	66,317 (37.9%)	104,728 (59.9%)	65,907 (37.7%)	84,039 (48.0%)	156,446 (89.4%)

Table 3: Label distribution per model and majority vote across all 174,973 tokens (count and % of total). LLaMA's NONE count includes 954 unparseable responses.

Gold label	Count	%
Lexical innovation	599	58.7
of which NEOLOGISM	465	45.5
of which ENTITY	134	13.1
Non-neologism	422	41.3
of which FOREIGN	61	6.0
of which NONE	361	35.4
Total	**1,021**	**100**

Table 4: Gold annotation of the 1,021 pipeline output candidates.

class taxonomy, following the annotation criteria derived from the theoretical framework in §3: tokens were classified as NEOLOGISM if they resulted from a word-formation process (grammatical or extra-grammatical in the sense of Mattiello 2013) and were first attested after 2015, as ENTITY if they denoted a proper noun first attested after 2015, as FOREIGN if they belonged to another language, and as NONE otherwise. Table 4 cross-tabulates pipeline output against gold labels.

Of the 1,021 candidates, 599 (58.7%) are genuine lexical innovations: 465 neologisms and 134 named entities. The remaining 422 consist of 361 false positives (NONE) and 61 foreign-language tokens that escaped both rule-based and LLM-based detection.

7. Discussion

The pipeline is best understood as a high-recall candidate generator rather than a precision classifier. Its primary contribution is the 122,031:1 compression ratio, which reduces a task that no human annotator could feasibly undertake (reviewing 124.6 million tokens) to one that a single annotator can complete (reviewing 1,021 candidates). We do not report corpus-level recall, as the gold standard covers only the pipeline output; the number of neologisms in the 124.6 million tokens that the pipeline may have missed is unknown. To give a rough idea, however, an estimate based on an external reference list is provided in §7.3. This framing aligns with how comparable systems are evaluated: Tomaszewska et al. (2025) report precision at each stage and note that recall is not computable; Grieve et al. (2018) and Brasolin et al. (2023) similarly report counts of emerging words found. As in those systems, a final manual verification step is an integral part of the design, not a limitation.

Process	Examples
Analogical formations	
Extra-gramm.: surface analogy	*updoot, pawrents*
Extra-gramm.: analogy via schema	
Secreted c.f.: *-fluencer*	*finfluencer, fitfluencer*
Abbreviated c.f.: *trad-*	*tradwife, tradfem*
Non-analogical formations	
Grammatical: prefixation	*deplatform, exvegan*
Grammatical: suffixation	*wokeism, trumpism*
Grammatical: compounding	*deepfake, longcovid*
Marginal: neoclassical c.f.	*abrosexual, acephobia*
Extra-gramm.: blending	*barbenheimer, maskne*
Extra-gramm.: expressive morph.	*thiccest, consoomer*

Table 5: Word-formation processes among gold neologisms.

7.1. Word-Formation Patterns

The 599 gold lexical innovations exhibit a range of word-formation processes that connect directly to the theoretical frameworks in §3. Table 5 organises the most productive patterns along two axes: whether the formation is analogical or non-analogical, and whether the process is grammatical, marginal, or extra-grammatical in the sense of Mattiello (2013).

Among non-analogical formations, standard grammatical processes account for a substantial share of the data. Prefixation with productive English prefixes yields forms such as *deplatform*, *detrash*, *exvangelical*, and *exvegan*. Suffixation with *-ism* generates *wokeism*, *trumpism*, *defaultism*, and *longtermism*. Compounding produces *deepfake*, *longcovid*, and *vibecheck*. At the margins of grammatical morphology, neoclassical combining forms of Latin or Greek origin appear in novel bases: *-sexual* (*abrosexual*, *dreamsexual*) and *-phobia/-phobic* (*acephobia*, *enbyphobic*).

Non-analogical extra-grammatical processes include blending, where two source words are fused without following a prior model (*barbenheimer*, *maskne*, *trumpanzee*), and expressive morphology, where deliberate phonological distortion of existing words produces new forms (*thiccest*, *consoomer*, *chonkster*).

The analogical formations are exclusively extra-grammatical. Surface analogy, where a single word serves as model, accounts for cases such as *updoot* (after *upvote*) and *pawrents* (after *parents*). More productive are formations arising through analogy via schema, where a recurrent fragment extracted from an initial blend becomes a combining form used across a series. Several such combining forms have undergone semantic generalisation, or secretion in the terminology of Mattiello (2013): *-fluencer* (from *influencer*; *finfluencer*, *fitfluencer*, *scamfluencers*), *-cel* (from *incel*; *femcel*, *mentalcels*), *-core* (from *hardcore*; *goblincore*, *traumacore*), *-nomics* (from *economics*; *bidenomics*, *tokenomics*), *-pilled* (from *redpilled*; *blackpilled*, *blackpillers*), and *-maxxing* (from *maxing*; *looksmaxxing*, *gymmaxxing*). Others function as abbreviated combining forms without semantic reinterpretation: *trad-* (from *traditional*; *tradwife*, *tradfem*) and *-flation* (from *inflation*; *greedflation*, *pissflation*). Whether the secreted forms have fully detached from their source words or remain at an intermediate stage between splinter and combining form is a diachronic question that the present data cannot resolve.

7.2. Error Analysis

The 422 non-neologistic tokens in the output fall into distinct categories, each pointing to a specific pipeline limitation.

False positives (361 tokens). The largest error category comprises concatenations (two or more words typed without a space, a common Reddit orthographic artifact) such as *datingapp*, *sidehustle*, and *telegramchannel* (approximately 50 tokens). These were correctly identified and removed by the word segmentation step, but subsequently reintegrated by the frequency threshold mechanism (§4.5), which restores all tokens with $\geq$100 occurrences regardless of the reason for their exclusion. At that point, the LLMs should have classified them as NONE, but failed to recognise them as mere orthographic variants of existing word sequences. A second cluster (approximately 50 tokens) comprises tokens from gaming and technical domains marked NONE for heterogeneous reasons. Some are names of programming functions or UI components (*floatlayout*, *floatmenu*, *floattensor*): accepting these would entail treating entire programming language vocabularies as natural language neologisms. Others are spaceless concatenations of pre-existing proper names (*bioniccommando*, a 1987 Capcom title; *biorepeel*, a cosmetic brand). Still others are fragments of game-internal proper names where the pipeline captured only part of a multi-token entity, or tokens that predate 2015 but were too domain-specific for the reference vocabulary. Approximately 30 tokens are misspellings of neologisms themselves (*neurodivegent*, *dollification*, *nuerotypicals*): the typo filter checks edit distance against *dictionary* words only and cannot detect that *neurodivegent* is a misspelling of *neurodivergent*, which is itself absent from the reference vocabulary. Finally, approximately 25 tokens are pre-2015 words absent from the 16.3 million-word reference vocabulary (*latinx*, *onfleek*, *biliteracy*): the vocabulary is comprehensive but not exhaustive for informal and slang registers.

Foreign language leakage (61 tokens). Two patterns dominate. Taglish (Tagalog–English) code-switching accounts for 29 of 61 tokens (47.5%): Tagalog prefixes (*na-*, *naka-*, *sina-*) affixed to English roots (*naghost*, *nakablock*, *sinasuggest*) superficially resemble English words with unfamiliar morphology, evading both Lingua and LLM classification. The remaining cases are English loanwords with Romance or Germanic inflection (*influenciador*, *influenceuse*, *brunchen*, *stressar*), originating from non-English posts where the mixed morphology falls below the language detector's confidence threshold.

The neologism–entity boundary. The 134 named entities in the gold standard highlight an inherently fuzzy boundary. Tokens such as *superstonk* (the subreddit name that became synonymous with the GameStop movement) and *barbenheimer* (*Barbie* + *Oppenheimer*) denote specific referents while also exhibiting productive word-formation processes (compounding, blending), making the neologism–entity distinction a matter of annotation judgment rather than a clear-cut category. Game-specific terms from *Splatoon* (Nintendo, 2015), numbering 15 tokens, *Among Us*, and various crypto projects account for the bulk of entities, and many are concentrated in one or two subreddits, a pattern that a cross-subreddit dispersion threshold could help address (see the Limitations section).

7.3. Pipeline False Negatives

To estimate recall, we compile a reference list of 103 single-token neologisms documented after 2015 in major dictionaries and lexicographic sources: Merriam-Webster additions (2016–2025),

Oxford and Collins Words of the Year, the American Dialect Society Word of the Year, the British Council "90 Words" list, Cambridge Dictionary, the OED, and Wiktionary, as well as community-maintained documentation sources such as Know Your Meme (full list in Appendix B). Of the 103 reference items, 20 are correctly detected; 48 were already attested on Reddit before 2015 and are correctly excluded; and 2 are excluded from evaluation as inflected forms of detected base forms. Loanwords (e.g. *mukbang* from Korean) are excluded, since borrowing falls outside the scope of the word-formation frameworks adopted in §3 (see the Limitations section). Inflected forms of base forms already detected by the pipeline (e.g. *deepfakes* alongside *deepfake*) are likewise excluded from the false negative count, as the pipeline's purpose is to identify novel lexical items rather than to capture every inflectional variant. The 33 genuine false negatives fall into the following categories: vocabulary homograph conflicts (17 tokens), where sparse pre-2015 occurrences in unrelated senses block the neologism (e.g. *rizz* as a character name, *simp* as a gaming clan); external vocabulary matches (9 tokens), where WordNet, Wiktionary, or Wikipedia contain the word under a different meaning (*copium*, *doggo*, *stonks*), a problem closely related to the semantic shift limitation discussed below; concatenation detection (3 tokens), where the segmentation module splits the token into known substrings (e.g. *cottagecore*); and the remaining 4 tokens are lost to tokenisation, typo correction, or LLM misclassification. Note that *doggo*, used in §3.2 as a canonical example of extra-grammatical morphology, is missed precisely because Wiktionary lists it under its pre-existing adverbial sense. Recall over the 53 genuinely post-2015 items is 20/53 (37.7%). Conversely, *vtuber* (virtual YouTuber), initially flagged as a typo of *tuber* by SymSpell, was correctly reintegrated by the frequency threshold mechanism owing to its 41,024 occurrences, illustrating that the reintegration stage (§4.5) functions as an effective safety net for high-frequency neologisms. The main axis for improvement is therefore conservative refinement of the vocabulary filtering stage, where type-level matching without sense disambiguation remains the primary source of false negatives.

8. Conclusion

We presented a scalable pipeline for automatic neologism detection that combines rule-based filtering with multi-model LLM classification, grounded in grammatical and extra-grammatical word-formation theory. Applied to 527 million Reddit posts, the pipeline achieves a 122,031:1 compression ratio, yielding 1,021 candidates of which 599 (58.7%) are genuine lexical innovations. Manual analysis of the output reveals a range of productive word-formation processes, from standard prefixation and compounding to analogical patterns such as secreted combining forms, confirming that the pipeline captures theoretically meaningful variation. The error analysis indicates that future improvements should target the earliest filtering stages rather than downstream classification. The pipeline code, vocabulary compilation scripts, and the annotated candidate list are available at https://github.com/DiegoRossini/neologism-pipeline.

Ethics Statement

The pipeline operates on unmoderated social media data and the resulting candidate list inevitably contains tokens related to offensive language, sexual content, hate speech, and extremist ideologies. Their inclusion reflects the lexical productivity of these domains and does not imply endorsement. All data was processed at the token level; no individual users were identified or tracked.

Limitations

The pipeline detects only single-token neologisms. Multi-word expressions such as *rage bait* or *touch grass* are invisible to the current architecture because the tokenizer treats each word independently, and no downstream stage attempts to reassemble multi-token units. Multi-word expressions are a major vector for lexical innovation in informal registers, and their absence from the output means the pipeline systematically underrepresents phrasal coinages.

Neologisms containing numerals or non-alphabetic characters are likewise excluded: the pattern cleaning stage (§4.3) discards all non-purely-alphabetic tokens. This filters out an increasingly productive category of lexical innovation known as algospeak (Steen et al., 2023; Aleksic, 2025), where users deliberately substitute letters with numbers or symbols to evade algorithmic content moderation (e.g., *$3X* for *sex*, *$trippers* for *strippers*), as well as named entities whose orthography includes digits, such as *4chan*.

The pipeline also cannot detect semantic shifts, where an existing word acquires a new meaning without any change in form. A prominent example is *Karen*, a conventional given name that underwent pejoration on Reddit and Black Twitter during the mid-2010s to denote an entitled, privileged white woman who weaponizes her social position. Because *Karen* is already present in the reference vocabulary as a proper noun, the pipeline excludes it at the vocabulary filtering stage, and no subsequent stage is equipped to detect that its usage distribution has changed.

The current instantiation targets English only. While the architecture is modular and transferable, all filtering resources, phonotactic rules, and frequency dictionaries are English-specific, and the LLM prompts are written in English. Adapting the pipeline to other languages would require substituting these components and re-evaluating the filtering thresholds. For morphologically rich languages with extensive inflectional paradigms, surface-form vocabulary matching may require either substantially larger observed-form vocabularies or an additional lemmatization step, though lemmatizing neologisms is itself problematic, since a lemmatizer trained on existing vocabulary may not reliably reduce novel forms to their base.

The foreign language detection stage (§4.6) cannot distinguish non-English corpus noise from genuine loanwords entering English. Lexical borrowing is not a word-formation process in either framework adopted in §3 — neither Štekauer (2002) nor Mattiello (2013) include it in their taxonomies, consistent with the position that borrowing and word formation are fundamentally distinct, though interacting, domains (ten Hacken and Panocová, 2020). This means that nativised loanwords such as *mukbang* or *hygge*, which are established terms in English, fall outside the pipeline's scope and are excluded either by the language detector or by the reference vocabulary.

As discussed in §7.3, the most consequential false negatives originate at the tokenization and vocabulary filtering stages, where tokens attested in pre-2015 data as probable typos (*stonk*, *monke*) or present in encyclopaedic sources (*copium*) are silently treated as known vocabulary and never enter the candidate pool. Two targeted improvements would address recurrent false positive patterns identified in §7.2. A cross-subreddit dispersion threshold would complement the raw frequency threshold by requiring candidates to appear across a minimum number of distinct subreddits, filtering concatenations such as *datingapp* that accumulate high frequencies within a single community through repeated orthographic error rather than deliberate coinage. A post-classification deduplication step comparing candidates by edit distance would catch misspellings of neologisms already captured by the pipeline (e.g., *neurodivegent* alongside *neurodivergent*).

All filtering thresholds (e.g. token length, edit distance, frequency, language detection confidence) were set based on preliminary experimentation rather than systematically optimised on a development set, as the computational cost of a full pipeline run (50–65 hours) makes exhaustive parameter search impractical. A systematic sensitivity analysis is left for future work.

Data and Code Availability

The pipeline code, vocabulary compilation scripts, and the annotated candidate list are available at https://github.com/DiegoRossini/neologism-pipeline.

Acknowledgements

This research was funded by the NCCR Evolving Language, Swiss National Science Foundation Agreement No. 51NF40_225146. We thank the anonymous reviewers for their helpful feedback, Ryan of NoSlang.com for generously sharing the abbreviation list, and Quirin Würschinger for personal communication regarding the NeoCrawler.

9. Bibliographical References

Adam Aleksic. 2025. *Algospeak: How Social Media Is Transforming the Future of Language*. Knopf, New York.

Sabine Arndt-Lappe. 2015. Word-formation and analogy. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation: An International Handbook of the Languages of Europe*, volume 2, pages 822–841. De Gruyter Mouton, Berlin.

Geert Booij. 2010. *Construction Morphology*. Oxford University Press, Oxford.

Paolo Brasolin, Greta H. Franzini, and Stefania Spina. 2023. "Ti blocco perché sei un trollazzo": Lexical Innovation in Contemporary Italian in a Large Twitter Corpus. *Journal of Italian Linguistics*, 35(2):123–145.

Maria Teresa Cabré and Lluís de Yzaguirre. 1995. Stratégie pour la détection semi-automatique des néologismes de presse. *TTR: Traduction, Terminologie, Rédaction*, 8(2):89–100.

Emmanuel Cartier. 2017. Neoveille, a web platform for neologism tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 95–98, Valencia, Spain. Association for Computational Linguistics.

Wolfgang U. Dressler. 2000. Extragrammatical vs. marginal morphology. In Ursula Doleschal and Anna M. Thornton, editors, *Extragrammatical and Marginal Morphology*, number 12 in LINCOM Studies in Theoretical Linguistics, pages 1–10. Lincom Europa, München.

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in French newspapers. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4337–4344, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jack Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics*, 46(4):293–319.

Louis Guilbert. 1975. *La créativité lexicale*. Langue et Langage. Larousse, Paris.

Peter Hohenhaus. 1998. Non-lexicalizability as a characteristic feature of nonce word-formation in English and German. *Lexicology*, 4(2):237–280.

Daphné Kerremans, Jelena Prokić, Quirin Würschinger, and Hans-Jörg Schmid. 2018. Using data-mining to identify and study patterns in lexical innovation on the web: The *NeoCrawler*. *Pragmatics & Cognition*, 25(1):174–200.

Daphné Kerremans, Susanne Stegmayr, and Hans-Jörg Schmid. 2012. The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change. In *Current Methods in Historical Semantics*, pages 59–96.

Annette Klosa-Kückelhaus and Harald Lüngen. 2018. New German words: Detection, description, and dictionary entry. In *Lexicography in the Digital Age*, pages 559–569. Euralex.

Lívia Körtvélyessy, Pavol Štekauer, and Pavol Kačmár. 2021. On the role of creativity in the formation of new complex words. *Linguistics*, 59(4):1017–1055.

Lívia Körtvélyessy, Pavol Štekauer, and Pavol Kačmár. 2022. *Creativity in Word Formation and Word Interpretation: Creative Potential and Creative Performance*, 1 edition. Cambridge University Press.

Taylor Mahler. 2020. Lexical Emergence on Reddit. *Lexis – Journal in English Lexicology*, 16.

Elisa Mattiello. 2013. *Extra-Grammatical Morphology in English: Abbreviations, Blends, Reduplicatives, and Related Phenomena*. Number 82 in Topics in English Linguistics. De Gruyter Mouton, Berlin.

Elisa Mattiello. 2017. *Analogy in Word-formation: A Study of English Neologisms and Occasionalisms*. Number 309 in Trends in Linguistics. Studies and Monographs. De Gruyter Mouton, Berlin.

Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin.

Antoinette Renouf. 1993. A word in time: First findings from dynamic corpus investigation. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English Language Corpora: Design, Analysis and Exploitation*, pages 279–288. Rodopi, Amsterdam.

Alain Rey. 1976. Néologisme: un pseudo-concept? *Cahiers de Lexicologie*, 28(1):3–17.

Stefania Spina, Paolo Brasolin, and Greta H. Franzini. 2024. Detecting emerging vocabulary in a large corpus of Italian tweets. *Research in Corpus Linguistics*, 13(1):139–170.

Ella Steen, Kathryn Yurechko, and Daniel Klug. 2023. You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media + Society*, 9(3).

Pavol Štekauer. 2001. Fundamental principles of an onomasiological theory of English word-formation. *Onomasiology Online*, 2:1–42.

Pius ten Hacken and Renáta Panocová, editors. 2020. *The Interaction of Borrowing and Word Formation*. Edinburgh University Press, Edinburgh.

Aleksandra Tomaszewska, Dariusz Czerski, Bartosz Żuk, and Maciej Ogrodniczuk. 2025. NeoN: A Tool for Automated Detection, Linguistic and LLM-Driven Analysis of Neologisms in Polish. ArXiv:2505.15426 [cs].

Quirin Würschinger. 2021. Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. *Frontiers in Artificial Intelligence*, 4:648583.

Arnold M. Zwicky and Geoffrey K. Pullum. 1987. Plain morphology and expressive morphology. In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, pages 330–340.

Pavol Štekauer. 1998. *An Onomasiological Theory of English Word-Formation*, volume 46 of *Studies in Functional and Structural Linguistics*. John Benjamins Publishing Company, Amsterdam.

Pavol Štekauer. 2002. On the Theory of Neologisms and Nonce-formations. *Australian Journal of Linguistics*, 22(1):97–112.

Pavol Štekauer. 2005. Onomasiological Approach to Word-Formation. In Marcel Den Dikken, Liliane Haegeman, Joan Maling, Guglielmo Cinque, Carol Georgopoulos, Jane Grimshaw, Michael

Kenstowicz, Hilda Koopman, Howard Lasnik, Alec Marantz, John J. McCarthy, Ian Roberts, Pavol Štekauer, and Rochelle Lieber, editors, *Handbook of Word-Formation*, volume 64, pages 207–232. Springer Netherlands, Dordrecht. Series Title: Studies in Natural Language and Linguistic Theory.

10. Language Resource References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset.

Wolf Garbe. 2012. SymSpell: Symmetric delete spelling correction algorithm.

Princeton University. 2011. WordNet 3.1.

Peter M. Stahl. 2022. Lingua: The most accurate natural language detection library for python.

Urban Dictionary. 2025. Urban Dictionary entry database. Scraped and filtered to entries predating 2015.

Watchful1. 2025. Subreddit comments/submissions 2005-06 to 2024-12. Per-subreddit split of the Pushshift Reddit dumps. Available via Academic Torrents.

Wikimedia Foundation. 2015a. Wikipedia: English article titles dump. Dump dated 2015-01-01.

Wikimedia Foundation. 2015b. Wiktionary: English edition dump. Dump dated 2015-01-01.

A. Prompt Templates

Both prompts are used identically across all four models (Qwen 72B, LLaMA 70B, Mistral Large 123B, and Claude Haiku).

Multi-token prompt (primary pass, 10 tokens per call).

```
TASK: Classify each token
into ONE category.

ENTITY - Pure proper nouns only
  (real/fictional): people, characters,
  companies, brands, products, games,
  movies, places, apps
  Examples: elon, pikachu, google,
  iphone, fortnite, reddit, tokyo

NEOLOGISM - New English words, slang,
  OR words derived from proper nouns
  Examples: doomscrolling, ghosting,
  rizz, bussin, adulting, covidiot,
  youtuber, redditor, trumpian,
  instagrammable, uberize, googlable

FOREIGN - Non-English words
  Examples: além, anspielung,
  yapmyorum, además

NONE - Usernames, typos, programming
  terms, unclear words

CRITICAL RULES:
1. Derived forms are NEOLOGISM
   (youtuber -> NEOLOGISM,
    youtube -> ENTITY)
2. When uncertain, classify as NONE
3. Use the context and subreddit
   to understand usage

TOKENS:
TOKEN: <token_1>
  context_1 (r/<subreddit>): "<text>"
  context_2 (r/<subreddit>): "<text>"
  context_3 (r/<subreddit>): "<text>"
TOKEN: <token_2>
  context_1 (r/<subreddit>): "<text>"
  ...

OUTPUT:
One classification per line as
TOKEN:LABEL (ENTITY, NEOLOGISM,
FOREIGN, or NONE).
No explanations.
```

Single-token prompt (retry pass for failed tokens).

```
Classify this token into ONE
category: ENTITY, NEOLOGISM,
FOREIGN, or NONE.

ENTITY - Pure proper nouns only
  (real/fictional): people,
  characters, companies, brands,
  products, games, movies, places,
  apps
NEOLOGISM - New English words, slang,
  OR words derived from proper nouns
  (youtuber, trumpian,
   instagrammable)
FOREIGN - Non-English words
NONE - Usernames, typos, programming
  terms, unclear words

TOKEN: <token>
  context_1 (r/<subreddit>): "<text>"
  context_2 (r/<subreddit>): "<text>"
  context_3 (r/<subreddit>): "<text>"

Answer with ONLY the label:
<token>:LABEL
```

B. Recall Reference List

Table 6 and Table 7 list the 103 single-token neologisms used for the recall evaluation in §7.3.

Status labels. TP = detected by the pipeline (true positive); FN = genuine false negative (post-2015, missed by pipeline); pre-15 = correctly excluded (attested on Reddit before 2015); excl. = excluded from evaluation (inflected form of a detected base form).

Source abbreviations. MW = Merriam-Webster; KYM = Know Your Meme; BC90 = British Council 90 Words; ADS = American Dialect Society; Collins = Collins Dictionary; Cambridge = Cambridge Dictionary; Oxford WOTY = Oxford Word of the Year; UrbanDict = Urban Dictionary; Dictionary.com = Dictionary.com; Aesth. Wiki = Aesthetics Wiki.[2]

[2]Source base URLs: Merriam-Webster: https://www.merriam-webster.com; Know Your Meme: https://knowyourmeme.com; British Council 90 Words: https://www.britishcouncil.org; American Dialect Society: https://www.americandialect.org; Collins Dictionary: https://www.collinsdictionary.com; Cambridge Dictionary: https://dictionary.cambridge.org; Oxford Word of the Year: https://languages.oup.com/word-of-the-year; Urban Dictionary: https://www.urbandictionary.com; Dictionary.com: https://www.dictionary.com; Wiktionary: https://en.wiktionary.org; Aesthetics Wiki: https://aesthetics.fandom.com. The full reference list with per-word verification URLs is available in the project repository.

Word	Year	Source	Status
doomscroll	2020	MW 2023	TP
doomscrolling	2020	MW 2023	TP
deepfake	2017	MW 2023; BC90	TP
deepfakes	2017	MW 2023	excl.
finsta	2017	MW 2023	FN
edgelord	2016	MW 2023; BC90	pre-15
copypasta	2016	MW 2023	pre-15
clickbait	2015	MW 2018	pre-15
subtweet	2015	MW 2018	pre-15
doxing	2015	MW 2017	pre-15
doxxing	2015	MW 2023	pre-15
ghosting	2017	MW 2017	pre-15
catfishing	2015	MW 2023	pre-15
copium	2020	Collins; MW	FN
hopium	2020	Collins	pre-15
shitposting	2017	Wiktionary	pre-15
shitpost	2017	Wiktionary	pre-15
rizz	2023	MW 2023; Oxford WOTY 2023; BC90	FN
simp	2019	MW 2023	FN
simping	2019	MW 2025	FN
stan	2017	MW 2019	pre-15
stanning	2017	MW 2019	pre-15
sealioning	2017	Collins	pre-15
doggo	2017	MW 2023	FN
birb	2017	KYM	FN
chonk	2018	KYM	FN
chonky	2018	KYM	FN
poggers	2017	KYM	FN
stonks	2021	KYM	FN
thicc	2017	KYM	FN
updoot	2016	KYM	TP
yeet	2018	MW 2023	pre-15
yeeted	2018	MW 2023	FN
sussy	2021	KYM	FN
bussin	2021	MW 2023	pre-15
skibidi	2023	Cambridge 2025	FN
delulu	2023	Cambridge 2025	FN
uwu	2017	KYM	pre-15
smol	2016	KYM	FN
blorbo	2022	KYM	TP
enshittification	2023	ADS WOTY 2023	TP
enshittify	2023	Wiktionary	FN
touchgrass	2021	MW 2024	FN
blockchain	2016	MW 2018	pre-15
cryptocurrency	2017	MW 2018	pre-15
bitcoin	2016	MW 2016	pre-15
chatbot	2017	MW 2018	pre-15
ransomware	2017	MW 2018	pre-15
deepfaked	2019	Wiktionary	FN
vtuber	2020	Wiktionary	TP
hodl	2017	Wiktionary	pre-15
defi	2020	Wiktionary	pre-15

Table 6: Recall reference list (1/2). "Year" indicates when the source documented the word, not the year of coinage.

Word	Year	Source	Status
altcoin	2017	Wiktionary	pre-15
memecoin	2021	Wiktionary	pre-15
stablecoin	2020	Wiktionary	pre-15
rugpull	2021	Wiktionary	FN
rugpulled	2021	Wiktionary	FN
wokeism	2019	Wiktionary	TP
wokeness	2019	Wiktionary	FN
trumpism	2016	Wiktionary	TP
deplatform	2018	Wiktionary	TP
deplatformed	2018	Wiktionary	excl.
deplatforming	2018	Wiktionary	TP
mansplaining	2015	MW 2018	pre-15
manspreading	2015	MW 2016	pre-15
whataboutism	2017	MW 2019	pre-15
incel	2018	Collins WOTY 2018; BC90	pre-15
incels	2018	Collins WOTY 2018	pre-15
blackpill	2018	Wiktionary	FN
blackpilled	2018	Wiktionary	TP
redpilled	2016	Wiktionary	pre-15
breadcrumbing	2018	Wiktionary	TP
situationship	2022	Oxford WOTY 2023; BC90	pre-15
allyship	2018	MW 2019	pre-15
covidiot	2020	Collins	TP
quarantini	2020	Wiktionary	FN
longcovid	2020	Wiktionary	TP
superspreader	2020	MW 2020	FN
doomscroller	2020	MW 2023	TP
covfefe	2017	Wiktionary	TP
infodemic	2020	Wiktionary	FN
deadname	2018	MW 2023	FN
deadnaming	2018	MW 2023	FN
genderfluid	2016	MW 2018	pre-15
demisexual	2018	Wiktionary	pre-15
neurodivergent	2020	MW 2023	pre-15
adulting	2016	MW 2017; BC90	pre-15
cottagecore	2020	Dictionary.com	FN
goblincore	2020	Dictionary.com	TP
darkcore	2020	Aesth. Wiki	pre-15
tradwife	2020	Cambridge 2025	TP
sponcon	2018	Wiktionary	FN
finfluencer	2020	Wiktionary	TP
hygge	2016	Oxford WOTY 2016	pre-15
glamping	2016	MW 2016	pre-15
athleisure	2016	MW 2016	pre-15
shadowban	2022	MW 2024	pre-15
jawnz	2023	UrbanDict	pre-15
longhauler	2020	Wiktionary	FN
rawdogging	2024	Wiktionary	pre-15
brainrot	2024	Oxford WOTY 2024	FN
airdrop	2017	Wiktionary	pre-15
gaslighting	2016	MW WOTY 2022	pre-15

Table 7: Recall reference list (2/2). "Year" indicates when the source documented the word, not the year of coinage.

High Resource Bias in AI-Driven Neology: Structural Inequality in Lexical Innovation

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Large language models (LLMs) are increasingly deployed to detect, generate, and normalize neologisms across languages. While prior work has examined their capacity to model semantic change and handle temporal drift, insufficient attention has been paid to how training data asymmetries interact with probabilistic generation mechanisms to structure lexical innovation itself. This position paper argues that AI-driven neology is shaped by systematic high resource bias that privileges dominant languages in the production, stabilization, and dissemination of new lexical items. Drawing on sociolinguistics, language political economy, lexicography, and computational modeling theory, we formalize how distributional imbalance alters innovation likelihood across languages. We introduce a taxonomy of bias types specific to AI-mediated neology, present a probabilistic account of generative reinforcement loops, and illustrate these mechanisms using documented examples from English-Arabic and English-Icelandic language pairs. We derive empirically testable predictions, outline concrete experimental protocols for their validation, and propose mitigation strategies for lexicographers, language planners, and NLP researchers.

Keywords: neology, large language models, linguistic bias, lexical innovation, high-resource languages, low-resource languages, language resources equity

1. Introduction

Neologisms emerge within speech communities through innovation, uptake, and stabilization. Historically, lexical change has been modeled as a socially distributed process observable through corpus frequency and contextual shift (Lejeune and Cartier, 2017). With the rise of large language models (LLMs), a new infrastructural actor enters this process: generative systems capable of producing plausible lexical forms largely independent of direct community grounding. These systems do not merely reflect existing language use; they actively shape the probability landscape in which new words are formed, evaluated, and propagated.

Contemporary LLMs are trained on corpora in which English and a handful of dominant languages are massively overrepresented (Joshi et al., 2020). This imbalance is not merely representational; it fundamentally reshapes the geometry of embedding spaces, the density of contextual neighborhoods, and the probability mass assigned to candidate lexical forms. As a result, generative outputs systematically favor patterns from high resource languages, creating structural asymmetries in lexical innovation that extend beyond simple performance degradation in low resource settings.

Recent empirical work has begun to document these effects. Zheng et al. (2024) demonstrate that even a single neologism can reduce machine translation quality by up to 43%, with effects more pronounced for words of non-English origin. Ármannsson et al. (2025) show reduced accuracy in morphological well-formedness judgments for Icelandic compared to English baselines. The comprehensive survey by Al-Khalifa et al. (2025) documents persistent preferences for English-derived transliterations over indigenous Arabic derivations in technical domains.

This paper advances three central claims:

1. AI-driven neology is structured by global inequalities in linguistic capital (Bourdieu, 1991).

2. Generative architectures amplify innovation originating in dominant languages while marginalizing or normalizing innovation in low resource contexts.

3. Without corrective mechanisms, LLM-integrated lexicographic practice risks reinforcing structural linguistic inequality on a global scale.

Our contribution is explicitly a position paper: it is theoretical and analytical in nature. We formalize mechanisms and derive testable predictions rather than presenting new benchmarking experiments, complementing empirical work such as NEO-BENCH (Zheng et al., 2024), the Icelandic linguistic benchmark (Ármannsson et al., 2025), and recent surveys of Arabic LLMs (Al-Khalifa et al., 2025). By focusing specifically on lexical innovation, we extend broader discussions of LLM bias (Navigli et al., 2023) to the domain of language change and resource equity. Importantly, we also outline concrete experimental protocols that would enable future empirical validation of our claims, responding to the need for actionable research directions in this emerging area.

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 16–26
16 MAY 2026.

2. Conceptual Delimitation

2.1. From Data Imbalance to Innovation Asymmetry

Data imbalance across languages has been widely documented in the NLP literature (Joshi et al., 2020; Navigli et al., 2023). However, the present argument concerns a distinct and previously under-theorized phenomenon: *innovation likelihood asymmetry*. Most existing discussions of imbalance focus on degraded performance in low resource languages, such as reduced translation quality or higher perplexity when encountering rare forms. Here, the focus shifts to generative dynamics.

Neology involves modeling productive morphological processes, semantic extension, compounding creativity, and lexical blending. These processes depend critically on dense distributional representations. The relevant question is therefore not only whether a language is underrepresented in training data, but whether the density of its contextual embedding space supports probabilistically plausible lexical innovation. High resource bias in neology is not reducible to general data imbalance; it concerns how imbalance actively restructures the innovation space itself.

This distinction is crucial because even morphologically rich low resource languages may exhibit suppressed indigenous creativity when mediated by current LLMs. Icelandic, with its productive compounding system, and Arabic, with its root-and-pattern morphology, both show evidence of this suppression despite their structural complexity (Ármannsson et al., 2025; Al-Khalifa et al., 2025; Wiemerslage et al., 2022).

2.2. Borrowing Versus Algorithmic Amplification

Borrowing is a natural and historically ubiquitous linguistic process. Languages routinely adopt foreign lexical material in domains of technological and cultural change. The argument here does not pathologize borrowing. Instead, it distinguishes between *organic, contact-driven borrowing* shaped by sociocultural interaction and *algorithmically amplified borrowing* driven by probabilistic reinforcement within digital infrastructures.

The issue is one of structural acceleration and asymmetry. When generative systems systematically increase the visibility and probability of dominant-language innovations, they may distort the ecological balance between borrowing and indigenous derivation. Recent lexicographic analyses document this effect in real-world dictionary compilation pipelines (Poix and Shevchenko, 2025).

2.3. LLMs as Infrastructural Mediators

This paper does not claim that LLMs autonomously create language change. Human communities remain the ultimate agents of stabilization and uptake. However, LLMs function as powerful infrastructural mediators within contemporary socio-technical networks. Gillespie (2014) argues that algorithms embedded in digital platforms increasingly determine what information is considered relevant, shaping participation in public life through procedural logics. Algorithmic systems shape visibility, salience, and circulation of linguistic forms. In generative contexts, they additionally influence which lexical candidates are more likely to be produced and repeated at scale.

As Periti and Montanelli (2024) observe in their survey of lexical semantic change through LLMs, these models fundamentally alter how we can detect, interpret, and assess meaning change over time. LLMs participate in language change not as originators but as amplifiers and redistributors of innovation probability (Navigli et al., 2023).

3. Related Work

Research on bias in large language models has grown rapidly, primarily addressing social stereotypes, toxicity, and performance disparities across demographic groups (Navigli et al., 2023; Gallegos et al., 2024). A foundational contribution by Joshi et al. (2020) documented severe underrepresentation of the majority of the world's languages in both NLP corpora and conference publications, establishing the data imbalance that underlies the present argument. Their taxonomy classified languages into six resource categories, with the vast majority falling into the lowest tiers. This imbalance reflects broader patterns of linguistic hierarchy that sociolinguists have long documented (De Swaan, 2001; Blommaert, 2010).

The political economy of large-scale language modeling has attracted increasing critical attention. Bender et al. (2021), in their influential analysis of the risks associated with ever-larger language models, highlight how training data sourced predominantly from the web systematically underrepresents marginalized communities and linguistic minorities. Conneau et al. (2020) demonstrate that while cross-lingual transfer learning can benefit low resource languages, trade-offs emerge between positive transfer and capacity dilution as model coverage expands. Similarly, Xue et al. (2021) show that massively multilingual models exhibit significant performance disparities across languages despite their broad coverage.

Subsequent empirical studies have quantified how this imbalance propagates to model behavior. Zheng et al. (2024) introduced NEO-BENCH,

a benchmark specifically designed to test LLM robustness to neologisms. Their results show that model performance is nearly halved in machine translation when a single neologism is introduced. Critically, they found that LLMs are affected differently based on the linguistic origins of words, with non-English neologisms posing greater challenges.

For morphologically complex low resource languages, Ármannsson et al. (2025) created the first manually curated linguistic benchmark for Icelandic LLMs. Native-speaker evaluation revealed markedly reduced accuracy in well-formedness judgments and morphological productivity tests compared with English baselines. Similarly, the comprehensive survey by Al-Khalifa et al. (2025) of Arabic LLMs highlights persistent challenges in handling Arabic's rich morphological system and a preference for English-derived forms in technical domains.

The Arabic case is particularly well documented in terms of resource availability. Surveys of freely available Arabic corpora have repeatedly demonstrated the imbalance between the language's massive speaker population and its comparatively limited digital resource base (Zaghouani, 2014). While substantial efforts have been made to build dialectal resources across multiple Arab countries (Zaghouani and Charfi, 2018; Bouamor et al., 2018; Charfi et al., 2019), these remain small relative to English-language resources and are concentrated in specific domains such as social media and news, leaving technical and scientific domains particularly underrepresented. This gap in domain coverage is directly relevant to our argument about neological innovation, as it is precisely in technical domains that new terminology emerges most actively.

Theoretical work on morphological productivity provides essential grounding for understanding these patterns. Bauer (2001) offers a comprehensive treatment of productivity measurement, emphasizing the scalar nature of morphological processes and the role of frequency in determining productive potential. Bybee (1995) establishes theoretical connections between token frequency and morphological representation that inform our formalization. Hamilton et al. (2016) demonstrate that statistical laws govern semantic change, with word frequency playing a key role in determining rates of meaning evolution, findings directly relevant to modeling innovation probability.

Lexicographic perspectives on AI-generated language have emerged only recently. Poix and Shevchenko (2025), in their eLex 2025 contribution, explicitly discuss the challenge of distinguishing organically occurring neologisms from synthetic LLM outputs in corpus data. They warn that AI-generated text may artificially inflate hapax legomena and distort diachronic frequency trends, raising urgent questions about the authenticity of corpus-based lexicographic evidence.

Complementary work on morphological processing by Wiemerslage et al. (2022) demonstrates that unsupervised paradigm completion for low resource languages remains fundamentally limited by sparse training signals. This limitation directly affects the capacity of LLMs to model productive morphological innovation in these languages.

The sociotechnical dynamics of algorithmic mediation have been theorized by Gillespie (2014), who argues that algorithms function as relevance-determining systems that shape public knowledge and participation. This framework illuminates how LLMs, as generative algorithms, may restructure the landscape of lexical innovation by privileging certain forms over others.

No prior publication has synthesized these threads into a unified formal account of high resource bias specifically targeting lexical innovation. The present paper fills this gap while remaining grounded in verified, peer-reviewed findings.

4. Theoretical Foundations

4.1. Neology and Lexicalization

Neologisms emerge through five primary mechanisms: morphological derivation, compounding, semantic shift, blending, and borrowing. The productivity of these mechanisms varies both synchronically and diachronically (Bauer, 2001). Successful lexicalization further requires sustained frequency growth, semantic stabilization, and eventual institutional recognition (Lejeune and Cartier, 2017). Traditional corpus linguistics treats frequency trajectories as direct evidence of community uptake, a relationship now formalized through diachronic word embeddings that reveal statistical laws governing semantic change (Hamilton et al., 2016).

In AI-mediated environments, however, synthetic generation fundamentally complicates this evidentiary basis. LLM outputs can rapidly create the appearance of frequency without corresponding human adoption (Zheng et al., 2024; Poix and Shevchenko, 2025). This raises urgent questions for lexicographers and language resource curators: How can we distinguish genuine community innovation from algorithmically amplified forms? What new methodologies are needed to track authentic lexical change in corpora increasingly contaminated by AI-generated text?

4.2. Linguistic Capital and Global Hierarchy

Language operates within a global hierarchy of symbolic power (Bourdieu, 1991). De Swaan (2001)

formalizes this hierarchy as a system in which languages occupy positions ranging from peripheral to supercentral, with English functioning as the hypercentral language connecting all others. Dominant languages accumulate institutional infrastructure, technological embedding, and cultural capital. Digital textual production mirrors and amplifies this hierarchy, a pattern that Blommaert (2010) characterizes as linguistic stratification within globalized communication systems.

Training corpora for today's LLMs are heavily skewed toward English and a small set of other high resource languages (Joshi et al., 2020). Bender et al. (2021) argue that such scale-driven approaches systematically underrepresent the linguistic diversity of the world's population, with consequences for both equity and quality. Cross-lingual representation learning demonstrates clear tradeoffs between positive transfer and capacity dilution as the number of languages increases (Conneau et al., 2020). This asymmetry is not static; it becomes operationalized in probabilistic generation. As Navigli et al. (2023) document, data selection bias in training corpora cascades into multiple forms of social and linguistic bias in model outputs. For neology specifically, this creates self-reinforcing loops that systematically disadvantage lexical creativity in low resource contexts.

4.3. Distributional Semantics and Innovation Space

The capacity of LLMs to model productive word formation depends on the density and quality of distributional representations. In high resource languages, dense contextual neighborhoods enable robust generalization to novel forms. Models can accurately predict which morphological combinations are plausible, which semantic extensions are natural, and which compounds are well-formed.

In low resource languages, sparse representations constrain these capacities. As Wiemerslage et al. (2022) demonstrate, morphological processing quality correlates strongly with training data availability. The implication for neology is that even when low resource languages possess rich productive morphological systems, LLMs may fail to model their creative potential accurately.

5. Formalizing High Resource Bias

5.1. Setup

Let L denote a language, D_L the effective training corpus size in tokens, and $P(w \mid c, L)$ the conditional token probability given context c. For high resource languages, $D_{\text{high}} \gg D_{\text{low}}$. This disparity yields more accurate estimation of conditional probabilities, denser contextual neighborhoods, and more robust modeling of morphological productivity.

5.2. Innovation Probability

Let n be a candidate neologism constructed via productive morphological processes. In a generative model,

$$P(n \mid c, L) \propto \exp(f_\theta(n, c, L))$$

where f_θ is the learned scoring function. For structurally parallel innovations across languages,

$$\mathbb{E}[P(n \mid c, L_{\text{high}})] > \mathbb{E}[P(n \mid c, L_{\text{low}})]$$

because subword representations are better optimized, productive patterns are observed at higher frequency, and contextual embeddings exhibit substantially lower uncertainty.

We note that this formalization is intended as an illustrative abstraction rather than a validated model. Its purpose is to provide a structured framework for generating testable hypotheses, which we detail in Section 8. The mathematical formulation captures the core intuition that data asymmetry translates into innovation asymmetry, and it is deliberately kept simple to highlight this relationship clearly rather than to model all relevant variables.

5.3. Morphological Productivity

Let M_L represent the modeled productivity of morphological transformations. Following Bauer (2001) and Bybee (1995) on the relationship between frequency and morphological productivity, we hypothesize:

$$M_L \propto \log(D_L)$$

As corpus size increases, the model's capacity to generalize productive transformations grows nonlinearly. Consequently, low resource languages exhibit more conservative generation behavior, reduced rates of indigenous derivation, and greater reliance on high-frequency borrowed tokens. This pattern is consistently observed in both Icelandic compounding (Ármannsson et al., 2025) and Arabic morphological systems (Al-Khalifa et al., 2025).

5.4. Generative Reinforcement Dynamics

Let $P_t(n, L)$ denote the probability of generating neologism n at time t. If n originates in a high resource language,

$$P_{t+1}(n, L_{\text{high}}) = P_t(n, L_{\text{high}}) + \alpha \cdot \text{Exposure}_t$$

Through global digital circulation, the form gains visibility. In low resource languages the update becomes

$$P_{t+1}(n, L_{\text{low}}) = P_t(n, L_{\text{low}}) + \beta \cdot \text{TranslationExposure}_t$$

where typically $\beta > \alpha$ for borrowed forms due to their higher baseline probability in the model.

This produces a feedback loop: high resource innovation $\rightarrow$ AI generation $\rightarrow$ digital uptake $\rightarrow$ corpus reintegration $\rightarrow$ increased generation probability. Such loops accelerate linguistic homogenization, as documented in broader analyses of LLM bias propagation (Navigli et al., 2023).

6. Taxonomy of Bias Types

We identify four interlocking bias types specific to AI-mediated neology:

Type 1: Distributional Bias. Unequal modeling quality resulting from corpus size disparities produces sparser representations for low resource languages (Joshi et al., 2020). This bias affects the foundational capacity to represent and manipulate lexical forms.

Type 2: Generative Amplification Bias. Disproportionate reproduction and probability boosting of dominant-language innovations during generation (Zheng et al., 2024). High resource neologisms receive higher generation probabilities even when low resource alternatives exist.

Type 3: Translational Normalization Bias. Flattening of indigenous semantic nuance when LLMs default to high resource lexical templates during translation or cross-lingual tasks. This is particularly evident in Arabic technical neology, where transliteration often supersedes productive root-and-pattern derivation (Al-Khalifa et al., 2025).

Type 4: Institutional Adoption Bias. Faster validation and lexicographic acceptance of high-visibility generative forms. AI-amplified neologisms may achieve apparent frequency thresholds for dictionary inclusion more rapidly, complicating the detection of organic innovation (Poix and Shevchenko, 2025).

These biases interact multiplicatively, producing compound effects on global lexical ecosystems. A neologism disadvantaged by distributional bias will also suffer reduced generative amplification, face stronger normalization pressure toward dominant-language equivalents, and experience slower institutional recognition.

7. Illustrative Case Studies

Table 1 summarizes selected neologisms drawn from published benchmarks. All observations are based on empirical results reported in the cited literature. While we do not introduce new experimental analyses here, these examples serve to ground our theoretical framework in documented findings, illustrating how the bias types identified in Section 6 manifest in practice across different language pairs.

7.1. Cross-Linguistic Patterns

Beyond the specific Arabic and Icelandic cases, broader patterns emerge from the empirical literature. Zheng et al. (2024) note that neologisms of different linguistic origins pose varying challenges: words borrowed into English from other languages (such as *pig butchering* from Mandarin) show compartmentalized understanding, while native English formations are more robustly represented. This asymmetry suggests that even within high resource English, the provenance of neologisms matters.

The COVID-19 pandemic provided a natural experiment in cross-linguistic neological dynamics. Technical terms like *coronavirus*, *lockdown*, and *social distancing* required rapid adaptation across languages. Observations suggest that high resource languages integrated these terms quickly and diversely, generating multiple synonyms and stylistic variants. Low resource languages, by contrast, showed slower integration and greater reliance on direct borrowing rather than calquing or indigenous derivation.

These patterns support our central claim: the generative dynamics of LLMs systematically favor high resource language innovation while constraining creativity in low resource contexts. The effects compound across the taxonomy we propose: distributional bias creates unequal starting conditions, generative amplification widens the gap, translational normalization flattens alternatives, and institutional adoption bias cements the outcomes.

7.2. English-Arabic Case Study

The English blend *doomscrolling* is densely represented in training data. NEO-BENCH demonstrates strong performance on definition tasks in high resource settings but dramatic degradation in machine translation (Zheng et al., 2024). Arabic exhibits a rich system of root-and-pattern morphology that supports productive technical neology. However, in generative outputs, transliteration or descriptive calques consistently predominate over productive derivation (Al-Khalifa et al., 2025).

Arabic's morphological system offers multiple productive mechanisms for neological derivation. The root-and-pattern system allows creation of new words through established templates: for instance, the root *k-t-b* (related to writing) generates *kitāb* (book), *kātib* (writer), *maktaba* (library), and *maktūb* (written). This system could theoretically accommodate technical neologisms through analogical extension. Similarly, Arabic possesses productive compounding mechanisms and established patterns for arabicization of foreign terms.

Despite these resources, surveys of Arabic LLMs reveal systematic preferences for transliteration (Al-Khalifa et al., 2025; Darwish et al., 2021). Technical

Neologism	Formation Type	Resource Context	Observed LLM Behavior
doomscrolling	Morphological blend	High (English)	Lower perplexity; strong definition generation; MT quality drops 43% when introduced as unknown form
pig butchering	Semantic calque (from Mandarin)	High via English	Compartmentalized knowledge; literal translations predominate over idiomatic rendering
stablecoin	Technical compound	High (English)	Accurate definition generation; successful cross-lingual transfer to related high resource languages
Icelandic compounds (e.g., *sýkingarþreyta*)	Productive compounding	Low (Icelandic)	Reduced accuracy in well-formedness judgments; lower Wug-test performance vs. English baselines
Arabic technical terms (e.g., *metaverse* equivalents)	Translational borrowing	Low (Arabic)	Strong preference for transliteration over indigenous root-and-pattern derivation
COVID-related neologisms	Multi-type	Variable	High resource languages show rapid integration; low resource languages show delayed and less diverse adaptation

Table 1: Selected neologisms illustrating high resource bias patterns. All entries are derived from empirical findings in cited published benchmarks and surveys.

terms like *internet*, *computer*, and emerging vocabulary such as *metaverse* are frequently rendered as phonetic borrowings rather than morphologically integrated forms. This pattern reflects the higher prior probability assigned to borrowed forms in training data. Even when Arabic language academies have proposed indigenous alternatives, the distributional dominance of English in training corpora biases outputs toward transliteration.

The problem is compounded by the fact that existing Arabic corpora, while growing in volume, remain concentrated in certain domains and registers. Surveys of freely available Arabic corpora have documented persistent gaps in technical and scientific writing (Zaghouani, 2014), and while large-scale dialectal corpora now cover social media registers across multiple Arab countries (Zaghouani and Charfi, 2018; Bouamor et al., 2018; Charfi et al., 2019), technical neology remains poorly represented. This domain mismatch means that LLM training data for Arabic overrepresents informal registers where borrowing is already prevalent, further reinforcing the preference for transliterated forms over indigenous derivations.

The effect creates a self-reinforcing cycle: borrowed forms dominate corpora, models learn to prefer borrowed forms, generated text contains more borrowed forms, and future training corpora inherit this bias. This dynamic threatens the productivity of Arabic's morphological system in precisely the domains, such as technology and science, where neological creativity is most needed.

7.3. English-Icelandic Case Study

Icelandic language planning has long promoted indigenous coinages through institutions such as the Árni Magnússon Institute. The language possesses extraordinarily productive compounding and derivational systems. Icelandic has historically coined native terms for modern concepts: *sími* (telephone, from an old word for thread), *tölva* (computer, from *tala* 'number' + *völva* 'prophetess'), and *sjónvarp* (television, literally 'vision-throw'). This tradition reflects deliberate language policy aimed at maintaining linguistic purity and ensuring that Icelandic remains fully functional for expressing modern concepts.

Yet LLM outputs in hybrid prompts frequently default to English technical terms or hybrid forms (Ármannsson et al., 2025). The benchmark created by these researchers specifically tests morphological productivity through tasks including well-formedness judgments, Wug tests (requiring generation of novel inflected forms), and compound interpretation. Results show that state-of-the-art models perform significantly worse on Icelandic morphological tasks compared to structurally analogous tasks in English.

This disparity is particularly striking given Icelandic's morphological regularity. The language's inflectional system, while complex, follows highly predictable patterns that should, in principle, be learnable from sufficient data. The performance gap therefore reflects not inherent difficulty but training data distribution. With approximately 350,000 native speakers, Icelandic is dwarfed in corpus representation by English's billions of speakers and massive digital footprint.

The Icelandic case reveals a fundamental ten-

sion between probabilistic modeling and institutional language policy. When LLMs consistently suggest English borrowings over indigenous Icelandic compounds, they work against decades of careful language planning. Users interacting with AI systems may increasingly encounter and adopt these borrowed forms, potentially undermining the ecosystem of indigenous neological creativity that language planners have cultivated. This represents a concrete mechanism by which AI systems may accelerate language shift even in communities with strong institutional support for linguistic maintenance.

8. Empirically Testable Predictions

Based on our formalization, we derive three specific predictions amenable to empirical validation. For each prediction, we outline a concrete experimental protocol that would enable systematic testing, responding to the need for actionable research designs that can move the field from theoretical argument to empirical investigation.

Prediction 1: Generative Diversity Hypothesis. Under symmetric prompting conditions, high resource languages will exhibit significantly higher rates of indigenous morphological innovation than low resource languages. This can be tested via controlled generation experiments extending the methodology of Zheng et al. (2024), comparing neologism generation rates across typologically similar language pairs with different resource levels.

Proposed protocol: Design a set of parallel prompts in matched language pairs (e.g., English vs. Icelandic, English vs. Arabic) that elicit neologism generation for identical novel concepts. Using multiple LLMs (both proprietary and open-weight), collect at least 100 generated responses per language per model. Annotate each generated neologism for formation type (indigenous derivation, compounding, borrowing, calque, transliteration) using trained native speaker annotators. Compute the ratio of indigenous formations to borrowed forms across languages and test for statistically significant differences using appropriate non-parametric tests given the expected non-normal distributions.

Prediction 2: Borrowing Amplification Hypothesis. The probability of borrowed forms in low resource languages will increase measurably following global exposure of dominant-language neologisms. This is testable via temporal corpus analysis comparing borrowing rates before and after major LLM deployment waves, particularly for technical vocabulary domains.

Proposed protocol: Construct time-stamped corpora for Arabic and Icelandic technical writing spanning two periods: pre-ChatGPT (2018-2022) and post-ChatGPT (2023-2026). For each period, extract neologisms related to technology, AI, and digital culture. Measure the proportion of borrowings versus indigenous formations in each period. Control for the natural increase in borrowing by comparing rates in domains where LLM-generated text is prevalent (e.g., online content) versus domains where it is rare (e.g., print publications, academic writing). A significant increase in borrowing rates disproportionately concentrated in LLM-saturated domains would support this prediction.

Prediction 3: Morphological Suppression Hypothesis. AI outputs for morphologically rich low resource languages will show lower morphological novelty compared with matched human corpora. This prediction is directly testable against the benchmarks established by Ármannsson et al. (2025) for Icelandic and the Arabic evaluation frameworks surveyed by Al-Khalifa et al. (2025).

Proposed protocol: Compile a corpus of LLM-generated text and a matched corpus of human-authored text in Arabic and Icelandic across the same domains and time periods. Measure morphological diversity using type-token ratio of morphological patterns, hapax legomena rates for derivational and compound formations, and the proportion of productive use of native morphological templates. Compare these metrics between LLM-generated and human-authored subcorpora, testing whether LLM text shows significantly reduced morphological novelty. This approach builds directly on the morphological analysis tools used in the Icelandic benchmark (Ármannsson et al., 2025) and can leverage existing Arabic morphological analyzers such as those surveyed in Darwish et al. (2021).

9. Implications and Mitigation

9.1. For Lexicography

AI-assisted corpus monitoring tools must incorporate mechanisms to distinguish organic uptake from synthetic amplification. Lexicographic workflows should adopt generative provenance tracking, flagging items that may have entered corpora through AI generation rather than community usage (Poix and Shevchenko, 2025). This may require new metadata standards for corpus annotation and revised criteria for dictionary inclusion.

9.2. For Language Policy

Language planning institutions must explicitly account for algorithmic reinforcement of borrowing. Organizations such as the Árni Magnússon Institute and Arabic language academies face new challenges in promoting indigenous terminology when probabilistic systems systematically favor borrowed forms. Promising mitigation avenues include equity-

aware fine-tuning and retrieval-augmented generation grounded in carefully curated local corpora.

9.3. For NLP Research

Evaluation metrics should incorporate cross-lingual innovation parity rather than focusing solely on aggregate performance benchmarks (Joshi et al., 2020). Concrete mitigation strategies include:

1. Balanced multilingual pre-training with explicit low resource upsampling
2. Morphology-aware tokenization schemes tailored to low resource languages (Wiemerslage et al., 2022)
3. Community-in-the-loop validation pipelines for neologism detection
4. Development of neology-specific benchmarks for low resource languages extending existing dialectal and multi-genre corpus efforts (Bouamor et al., 2018; Charfi et al., 2019)

LREC is ideally positioned to lead by developing multilingual neology resources that explicitly tag generative versus human provenance.

10. Discussion

The mechanisms formalized in this paper suggest that current LLM architectures do not merely reflect existing linguistic inequalities; they actively accelerate them within digital ecosystems. Over time, this may lead to reduced lexical diversity worldwide, with low resource languages increasingly functioning as recipients rather than co-creators of neological innovation.

The Icelandic and Arabic case studies illustrate how even languages with strong institutional support and rich morphological systems remain vulnerable. Icelandic, with its centuries-long tradition of linguistic purism and active language planning, faces pressure from AI systems that consistently prefer English borrowings. Arabic, with over 400 million speakers and a morphological system of remarkable productivity, sees its derivational potential underutilized as models default to transliteration.

10.1. Broader Consequences for Linguistic Diversity

The implications extend beyond academic concern: lexical innovation is a core mechanism of cultural expression and adaptation. Languages evolve through their speakers' creative responses to new experiences, technologies, and social configurations. When AI systems systematically suppress indigenous creativity while amplifying borrowed forms, they may contribute to broader processes of cultural homogenization, a dynamic that parallels historical patterns of language endangerment (Crystal, 2000).

Consider the domain of technology, where neological activity is most intense. If speakers of low resource languages consistently encounter AI-generated text that favors English borrowings, they may internalize these preferences. The mT5 model, despite covering 101 languages, demonstrates clear performance disparities across resource levels that reflect underlying training data imbalances (Xue et al., 2021). Over generations, this could erode the productive capacity of morphological systems that require active use to remain vital. The result would be languages that retain their grammatical structures but increasingly rely on borrowed vocabulary for modern domains, a pattern historically associated with language shift and endangerment.

10.2. Implications for Language Documentation

For endangered and low resource languages, these dynamics pose particular challenges. Language documentation efforts increasingly rely on computational tools for corpus building, lexicographic work, and language learning materials. If these tools systematically underrepresent indigenous neological patterns, documentation may inadvertently encode a biased snapshot of the language. Future revitalization efforts would then inherit these biases, potentially perpetuating reduced lexical creativity even in human-mediated contexts.

10.3. Toward Equity-Aware Language Technology

Future interdisciplinary collaboration between computational linguists, sociolinguists, lexicographers, and language communities will be essential to design equity-aware systems that preserve rather than erode global linguistic creativity. The LREC community, with its emphasis on language resources for all, is well-positioned to lead this effort. Concrete steps include developing neology-specific benchmarks for low resource languages, creating curated corpora of indigenous technical terminology, and establishing best practices for generative provenance tracking in lexicographic workflows.

10.4. Scope and Position of This Contribution

It is important to situate this paper clearly within the broader research landscape. The phenomena of

data imbalance and resource asymmetry across languages are well established in NLP (Joshi et al., 2020; Bender et al., 2021; Navigli et al., 2023). Our contribution does not claim novelty in identifying these asymmetries per se. Rather, the novelty lies in synthesizing these findings into a unified framework specifically targeting *lexical innovation*, a domain where the consequences of bias have distinct and under-explored implications for linguistic diversity, language policy, and lexicographic practice.

As a position paper, this work is designed to serve as a conceptual foundation and research agenda. The taxonomy of bias types (Section 6), the formalized predictions (Section 8), and the experimental protocols outlined therein are intended to catalyze empirical investigation. We believe that the theoretical groundwork presented here is a necessary prerequisite for principled experimental design in this area, and we invite the community to build upon it.

11. Limitations

This paper advances a theoretical and formal argument rather than presenting new empirical measurements. The probabilistic formalization and the derived hypotheses are intended as structured explanatory abstractions that guide future quantitative investigation rather than as validated models. Systematic cross-linguistic benchmarking, controlled prompting studies, and longitudinal corpus analyses will be required to validate or refine the proposed claims.

All arguments are grounded in established literature on linguistic inequality, distributional modeling, and lexical innovation. However, the paper does not provide direct experimental evidence demonstrating differential innovation likelihood across specific model architectures. The case studies draw entirely on previously published empirical results rather than new analyses, which means they illustrate rather than independently confirm the proposed framework. As such, the framework should be interpreted as a structured explanatory hypothesis rather than a definitive empirical conclusion.

We also make no claim of universal applicability across every LLM architecture, training regime, or future model generation. Differences in tokenization strategies, multilingual balancing techniques, or morphology-aware modeling may mitigate or exacerbate the effects described here. Additionally, the illustrative case studies focus on English-Arabic and English-Icelandic language pairs because they represent contrasting sociolinguistic and policy environments. Extension to other language families, especially typologically distant or endangered languages, may reveal additional bias patterns or countervailing dynamics not captured in the present analysis.

Finally, the formalization abstracts away from complex sociopolitical variables that shape language use in digital environments, including state policy, educational systems, platform moderation practices, and economic incentives. These factors interact with model design in ways that warrant dedicated interdisciplinary study.

12. Conclusion

AI-driven neology is shaped by structural asymmetries in global textual production. Through probabilistic generation and reinforcement loops, LLMs may amplify dominant-language innovation while marginalizing indigenous lexical creativity. This paper has formalized these dynamics through a taxonomy of four interlocking bias types, a probabilistic framework for innovation likelihood, and three empirically testable predictions accompanied by concrete experimental protocols. Addressing the challenge requires sustained interdisciplinary collaboration and the adoption of equity-aware design principles across the language resource pipeline.

Without deliberate intervention, generative systems risk accelerating linguistic homogenization within digital ecosystems. We call on the LREC community to operationalize the proposed taxonomy, test the derived predictions using the experimental protocols outlined in this paper, and develop concrete resources that safeguard lexical diversity for future generations.

13. Ethical Considerations

If unexamined, high resource bias in AI-driven neology may contribute to the reinforcement of existing linguistic hierarchies. Amplification of dominant-language innovation, coupled with the normalization of borrowing patterns, can accelerate processes of semantic convergence and marginalize culturally embedded lexical practices. Over time, this may contribute to diminished visibility of indigenous knowledge systems and reduced incentives for community-based lexical development.

Ethical language technology development therefore requires structural transparency and participatory governance. We advocate for full disclosure of training data composition, including language distribution and sources, to enable independent auditing of cross-linguistic representation. Open, community-governed language resources should be prioritized to ensure that local innovation is documented and accessible for both training and evaluation purposes.

Moreover, researchers working on low resource and endangered languages should be included as

equal partners in the design, evaluation, and deployment of language technologies. Collaborative models that center community expertise can help prevent extractive data practices and ensure that technological development aligns with local linguistic priorities.

Equity-aware language modeling is not only a technical objective but also an ethical commitment to sustaining global linguistic diversity in increasingly AI-mediated communication environments.

Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

14. Bibliographical References

Al-Khalifa, S., Durrani, N., Al-Khalifa, H., and Alam, F. (2025). The landscape of Arabic large language models. *Communications of the ACM*, 68(10):54–61.

Ármannsson, B., Ingimundarson, F. Á., and Sigurðsson, E. F. (2025). An Icelandic linguistic benchmark for large language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 37–47, Tallinn, Estonia. University of Tartu Library.

Bauer, L. (2001). *Morphological Productivity*. Cambridge University Press, Cambridge.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, New York, NY. Association for Computing Machinery.

Blommaert, J. (2010). *The Sociolinguistics of Globalization*. Cambridge University Press, Cambridge.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Bourdieu, P. (1991). *Language and Symbolic Power*. Harvard University Press, Cambridge, MA.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10(5):425–455.

Charfi, A., Zaghouani, W., Mehdi, S. H., and Mohamed, E. (2019). A fine-grained annotated multi-dialectal Arabic corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 198–204, Varna, Bulgaria. INCOMA Ltd.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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.

Crystal, D. (2000). *Language Death*. Cambridge University Press, Cambridge.

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., Bouamor, H., Bouzoubaa, K., Cavalli-Sforza, V., El-Beltagy, S. R., El-Hajj, W., Jarrar, M., and Mubarak, H. (2021). A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, 64(4):72–81.

De Swaan, A. (2001). *Words of the World: The Global Language System*. Polity Press, Cambridge.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Gillespie, T. (2014). The relevance of algorithms. In Gillespie, T., Boczkowski, P. J., and Foot, K. A., editors, *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–193. MIT Press, Cambridge, MA.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (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.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Lejeune, G. and Cartier, E. (2017). Character based pattern mining for neology detection. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 25–30, Copenhagen, Denmark. Association for Computational Linguistics.

Navigli, R., Conia, S., and Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Periti, F. and Montanelli, S. (2024). Lexical semantic change through large language models: A survey. *ACM Computing Surveys*, 56(11):1–38.

Poix, C. and Shevchenko, N. (2025). The challenge of AI-generated neology. In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference*, pages 318–331, Bled, Slovenia. Lexical Computing.

Wiemerslage, A., Silfverberg, M., Yang, C., McCarthy, A. D., Nicolai, G., Colunga, E., and Kann, K. (2022). Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zaghouani, W. (2014). Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC 2014*, Reykjavik, Iceland.

Zaghouani, W. and Charfi, A. (2018). AraP-Tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Zheng, J., Ritter, A., and Xu, W. (2024). NEOBENCH: 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.

Do LLMs Know What Luxembourgish Borrows? Probing Lexical Neology in Low-Resource Multilingual Models

Nina Hosseini-Kivanani[*1,2]
[1] University of Luxembourg, Luxembourg
[2] Radio Télévision Luxembourg (RTL), Luxembourg
nina.hosseinikivanani@ext.uni.lu

Abstract

Large language models (LLMs) are increasingly used for writing assistance in small contact languages, yet it is unclear whether they respect community norms around lexical borrowing and neology. We introduce LexNeo-Bench, a 3,050-instance token-level benchmark derived from LuxBorrow, a large-scale Luxembourgish news corpus, where target tokens are labelled as native or as French, German, or English borrowings. Using this benchmark, we probe three multilingual LLMs across 34 prompt settings on two tasks: borrowing type classification and a binary lexical-innovation proxy (borrowing versus native). Without external context, models perform only slightly above chance on borrowing classification, so we construct a linguistic knowledge graph that encodes donor language, morphological patterns, and lexical analogues, and inject instance-specific subgraphs into the prompt. Knowledge-graph prompts raise borrowing classification accuracy from 25 – 35% up to 71 – 81% and largely close the gap between small and large models, while leaving neology detection difficult and sensitive to few-shot design. Our results show that lexicon-aware prompting is highly beneficial for robust borrowing judgments in low-resource contact languages and that lexical resources can serve as structured context for LLM evaluation. This study was carried out within the ENEOLI COST Action and examines borrowing as a form of lexical innovation in multilingual Luxembourgish data.

Keywords: Luxembourgish, lexical borrowing, neology, large language models, knowledge graphs

1. Introduction

Neology, the creation and diffusion of new lexical items, has long been central to lexicography, corpus linguistics, and sociolinguistics. With the emergence of large language models (LLMs), neology enters a new phase. LLMs are trained on massive multilingual corpora, absorb existing neologisms, and can themselves generate novel forms, blends, and hybrid structures in response to prompts. This raises questions not only about how LLMs detect and represent lexical innovation, but also about how their behavior interacts with existing norms and resources in individual language communities (Wolfer and Klosa-Kückelhaus, 2023; Zheng et al., 2024).

For smaller languages such as Luxembourgish, lexical innovation is tightly intertwined with contact phenomena. Luxembourgish exists in a dense contact zone with German, French, and English. Much of its modern lexical growth is realized through borrowing and adaptation from these donor languages rather than through entirely endogenous coinages (Adda-Decker et al., 2008). In written media, especially professionally edited news, many emergent forms are morphologically or orthographically integrated into Luxembourgish, while others remain closer to code-switching (Lavergne et al., 2014). For downstream Natural Language Processing (NLP) tools and LLM-powered applications, it matters whether these items are recognized as legitimate Luxembourgish words or treated as errors, foreign insertions, or targets for normalization back to French or German.

Previous work on Luxembourgish borrowing introduced LuxBorrow (Hosseini-Kivanani and Philippy, 2026), a large-scale corpus of Radio Télévision Luxembourg (RTL) news (1999–2025) annotated with sentence-level language identification and token-level labels for native items, borrowings from French, German, and English, and code-switching. That study focused on contact linguistic patterns and diachrony, showing that Luxembourgish remains the matrix language in news, while lexical borrowing and code mixing are pervasive but low-intensity, with a rich inventory of morphological and orthographic adaptation patterns. However, LuxBorrow did not address how contemporary LLMs treat these adapted forms, nor whether they recognize them as part of the Luxembourgish lexicon.

In this paper, we treat morphologically and orthographically adapted borrowings in Luxembourgish news as a key locus of lexical innovation and use LuxBorrow as ground truth to evaluate neology awareness in multilingual LLMs. We construct a token-level classification benchmark that pairs Luxembourgish sentences from RTL.lu with highlighted target tokens and gold labels indicating whether each token is native or a borrowing, and if so, from which donor language. On top of this benchmark, we define two tasks: a borrowing classification task in which models choose from

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 27–38
16 MAY 2026.

four labels (NATIVE, FR_LOAN, DE_LOAN, and EN_LOAN as a diagnostic distractor) but are evaluated on three gold classes, and a binary neology decision task.

Our study is organized around three research questions.

- RQ1. To what extent do off-the-shelf multilingual LLMs correctly classify native Luxembourgish words and distinguish French- vs German-origin adapted borrowings in RTL news?
- RQ2. Do LLMs systematically bias their judgments toward dominant donor languages, especially French and German, and how often do they incorrectly project English-origin hypotheses via the EN_LOAN distractor label?
- RQ3. How does providing explicit lexicon-based context, for example, a loanword registry, affect LLM performance and their treatment of Luxembourgish lexical innovation?

To answer these questions, we evaluate three strong multilingual LLMs in frozen, prompt-only mode. We compare zero-shot prompting, few-shot prompting with manually chosen examples of Luxembourgish borrowings, and two knowledge-based prompting conditions: `KG_flat`, which provides a global list of borrowing patterns, and `KG_graph`, which injects an instance-specific lexicon context derived from the LuxBorrow loanword registry.

Our contributions are threefold. First, we introduce LexNeo-Bench, a token-level benchmark for borrowing classification in Luxembourgish, derived from LuxBorrow, with public scripts for extraction, prompting, and evaluation. Second, we add a binary lexical-innovation proxy task that collapses borrowings versus native items to probe neology awareness, and show that it remains challenging even for strong multilingual LLMs. Third, we show that lightweight lexicon-based context via a linguistic knowledge graph can substantially improve borrowing judgments in a low-resource contact language, which suggests concrete avenues for integrating community-curated lexical resources into LLM prompting for neology-sensitive applications. Within the ENEOLI COST Action, this study contributes to WG2 by treating borrowing in Luxembourgish as a corpus-based case of lexical innovation and by evaluating how multilingual LLMs analyze such forms in a low-resource contact setting.

2. Related Work

2.1. Borrowing, code-switching, and neology

Contact linguistics distinguishes lexical borrowing, items integrated into the recipient language's lexicon and grammar, from code-switching, that is, spontaneous alternation between languages within discourse. Classic accounts emphasize that entrenched borrowings are morphologically and phonologically integrated, frequent, and often listed in dictionaries, while code-switches retain donor language structure and remain more speaker-specific. This view underlies the "Simple View" of borrowing, which operationalizes the difference in terms of listedness in the mental lexicon and community entrenchment (Treffers-Daller, 2025; Chesley and Baayen, 2010).

In multilingual European contexts, written media often show a stable matrix language with pervasive but shallow insertions from donor languages. Borrowings can be introduced via institutional domains such as politics, finance, and administration, before diffusing into more general registers. Over time, morphologically adapted forms may compete with native synonyms or with less integrated loan variants. This dynamic is particularly visible in Luxembourgish, where French and German both supply a rich inventory of technical and everyday lexical items, and where orthographic and morphological adaptation blur the surface boundary between native and borrowed forms (Anastasiou, 2022; Lavergne et al., 2014; Adda-Decker et al., 2008).

Lexicographic and corpus-based studies of neology therefore give prominence to borrowed and adapted items when tracking lexical innovation, especially in small languages that rely heavily on lexical importation from regional lingua francas (Wolfer and Klosa-Kückelhaus, 2023).

2.2. Computational borrowing and neology detection

In NLP, early work on multilingual text mixing emphasized document- or utterance-level indices, such as code-mixing indices and entropy-based measures, which treat all foreign tokens uniformly. More recent studies move to explicit borrowing detection and distinguish unassimilated foreign tokens, code-switches, and integrated loanwords. This line of work has introduced borrowing-annotated corpora, for example anglicism detection in Spanish newswire (Alvarez-Mellado, 2020, 2021), and shared tasks with sequence tagging baselines (Mellado et al., 2021; Álvarez-Mellado et al., 2025). Methods range from conditional random fields and BiLSTM-CRFs to

transformer taggers that incorporate lexical and orthographic features (Alvarez-Mellado, 2020; Álvarez Mellado, 2020), alongside resource-lean approaches to code-switching identification that rely mainly on word lists and monolingual corpora (Kevers, 2022).

Beyond borrowing per se, neology detection has traditionally relied on dictionary versus corpus comparisons combined with temporal information, for example, locating forms that appear in recent corpora but are absent from older lexica. With the advent of LLMs, recent work has begun to integrate these models into neologism detection pipelines, for example using them as filters or validators for candidate neologisms, and to provide lemmata and definitions for emergent forms (Tomaszewska et al., 2025; Hosseini-Kivanani, 2025). Other studies highlight how LLMs can also generate non-attested "LLM neologisms" due to tokenization and encoding artifacts (Iwamoto and Kanayama, 2024). This opens a new evaluation axis: not just whether LLMs can help detect neology, but whether their intrinsic lexical knowledge and biases align with community norms and lexicographic resources (Tomaszewska et al., 2025; Hosseini-Kivanani, 2025; Iwamoto and Kanayama, 2024).

2.3. LLMs, low resource languages, and lexical inequality

Work on LLMs in low-resource languages highlights skewed coverage and performance gaps, where models trained mainly on high-resource languages underrepresent or mis-analyze items from smaller languages and can "normalize" adapted borrowings back to donor forms. This has consequences for spell-checkers, assistive writing tools, and generation systems that interact with speakers of contact languages. Empirical studies of Luxembourgish resources and their multilingual context document sparse written production and heavy code-mixing and adaptation pressure, which exacerbate LLM coverage problems (Plum et al., 2024; Lavergne et al., 2014; Adda-Decker et al., 2008).

Lexicon-aware prompting and retrieval/gazetteer augmentation show that injecting compact community resources into LLM workflows yields large gains on complex Named Entity Recognition (NER) and entity-centric tasks (Tan et al., 2023; Chen et al., 2022). This motivates using curated loanword registries or structured knowledge-graph hints to probe LLM judgments about borrowed and adapted forms. Against this background, we use a borrowing-annotated Luxembourgish news corpus as a neology resource to build LexNeo-Bench, a benchmark that probes LLM lexical decisions in a dense contact setting.

3. Experiments

Figure 1 summarizes the overall evaluation pipeline, from LuxBorrow-derived benchmark construction and LKG retrieval to prompt assembly and multilingual LLM evaluation.

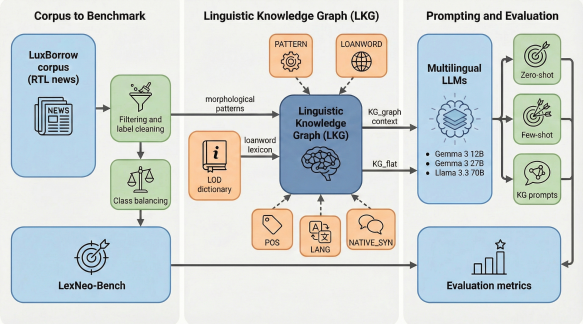

Figure 1: LexNeo-Bench pipeline.

3.1. Benchmark construction

We construct LexNeo-Bench, a token-level evaluation benchmark derived from the LuxBorrow corpus of professionally edited Luxembourgish news. LuxBorrow provides sentence-level language identification over RTL articles together with token-level borrowing labels generated by a morphological pattern pipeline. Each benchmark instance consists of a Luxembourgish sentence, a highlighted target token, its gold borrowing label, compact morphological evidence, and article metadata such as section and timestamp.

The label space follows the LuxBorrow taxonomy and distinguishes native Luxembourgish items (**NATIVE**) from French-, German-, and English-origin borrowings (**FR_LOAN**, **DE_LOAN**, **EN_LOAN**). Tokens tagged as **CODE_SWITCH** or as named entities are excluded from the main task, since the focus is on entrenched lexical items rather than span-level alternation. To avoid extremely sparse classes, we discard labels with fewer than 50 instances in the source corpus. EN_LOAN appears only 24 times and is therefore removed from the evaluation set, which yields a three-way task over NATIVE, FR_LOAN, and DE_LOAN, even though the conceptual four-way taxonomy is kept in the prompts.

The original LuxBorrow corpus comprises $259\,305$ RTL news articles and 43.7 million tokens. We first remove punctuation, a curated list of Luxembourgish function words, and tokens with low-confidence automatic labels. From the remaining pool, we draw 1000 instances per active class,

which results in a balanced benchmark of 3 000 examples, and we add a small diagnostic stratum of 50 CODE_SWITCH tokens for error analysis. The final benchmark therefore contains 3050 instances: NATIVE $= 1000$, FR_LOAN $= 1000$, DE_LOAN $= 1000$, and CODE_SWITCH $= 50$.

Each instance inherits the publication date of its source article (1999–2025), providing a diachronic signal. As a proxy for entrenchment, we contrast tokens from articles published before 2015 with those after 2015. This split is not first-attestation dating, but leverages a 25-year professionally edited news record. The borrowing labels serve as the primary gold standard for both tasks. For the auxiliary neology decision task, we derive a binary label at prompt construction time by mapping tokens annotated as `FR_LOAN`, `DE_LOAN`, or `EN_LOAN` to `YES` (lexical innovation) and `NATIVE` tokens to `NO`. The recent versus established flag is used only for temporal robustness analysis.

For prompt types that embed full lexicon entries, we restrict ourselves to a reduced subset of 674 benchmark items for which dictionary definitions, etymology, and related lexical information are available in a consistent format from the Lëtzebuerger Online Dictionnaire (LOD) (Zenter fir d'Lëtzebuerger Sprooch, 2025). This avoids noisy or incomplete context in lexicon-assisted conditions.

3.2. Linguistic Knowledge Graph

To provide models with structured linguistic context, we construct a **Linguistic Knowledge Graph (LKG)** that integrates three LuxBorrow-related resources: a compiled index of productive morphological patterns, a loanword lexicon extracted from LOD, and a hand-curated table of loanword–native synonym pairs.

Nodes are typed as morphological patterns (PATTERN), donor languages (LANG), loanword entries (LOANWORD), native Luxembourgish synonyms (NATIVE_SYN), and part-of-speech tags (POS). Edges encode linguistic relations such as pattern membership (*follows_pattern*), donor origin (*from_donor*), synonymy (*has_synonym*), competition between patterns with the same Luxembourgish affix but different donor languages (*contrastive*), and lexical category (*has_pos*).

For each benchmark instance, we perform multi-hop retrieval on this graph to construct a compact, token-specific explanation subgraph. Starting from the target token, we match compatible LuxBorrow patterns, query LOD to obtain donor metadata, and reconstruct an etymology-style chain of the form donor form $\rightarrow$ adaptation pattern $\rightarrow$ Luxembourgish form, collect linked native synonyms, and sample a small set of analogues that share a pattern, plus a few contrastive patterns with the same affix but different donor languages. The retrieved subgraph is then linearized into a structured natural-language block of at most 30 lines and prepended to the model prompt. This instance-specific `KG_graph` context replaces a much coarser `KG_flat` baseline in which the same global list of 19 patterns is appended to every example independently of the target token.

3.3. Prompt setups

All prompts share a common two-role template. The system message defines a Luxembourgish linguistics expert persona, and the user message concatenates optional knowledge-graph context, the task instruction, and the Luxembourgish sentence with the target token marked by `**`. The user message then introduces any external context, followed by a concise instruction to assign exactly one label to the highlighted token and the sentence in which it appears.

We evaluate a family of prompt strategies for both borrowing classification and neology detection. In all cases, the model receives a short English instruction, the Luxembourgish sentence, and the target token marked with `**`. For the **classification** task, the model must output exactly one label from the conceptual four-way set $\{$`NATIVE`, `FR_LOAN`, `DE_LOAN`, `EN_LOAN`$\}$. Although `EN_LOAN` does not appear in the evaluation data, we keep it as a possible answer to capture uncertainty toward English-origin candidates. For the neology task, the model must answer `YES` if the token should be treated as a lexical innovation in Luxembourgish, and `NO` otherwise.

The base prompt strategies include a plain `zero_shot` condition (system role plus task description, no additional context), a `few_shot` variant with five manually authored demonstrations, and a `minimal` variant that reduces the instruction to a single line and enforces label-only output. The five demonstrations do not overlap with the 3 050 benchmark instances. They consist of two prototypical `FR_LOAN` examples and one example each for `DE_LOAN`, `EN_LOAN`, and `NATIVE`. Each demonstration pairs a short Luxembourgish sentence with its gold label and a brief linguistically motivated justification.

An excerpt of the few-shot prompts is shown in Listing 1.

Listing 1: Excerpt of few-shot prompts for borrowing classification and neology decision.

```
System:
  You are a linguistic expert specializing in
     Luxembourgish
  (Lëtzebuergesch). Luxembourgish is a West
     Germanic language spoken
  in Luxembourg that regularly borrows and
     morphologically adapts
```

```
  words from French, German, and English.

User (classification task):
  Given the following Luxembourgish sentence and
     the highlighted
  word (marked with ** **), decide whether the
     highlighted word is:
    - NATIVE:  a native Luxembourgish word
    - FR_LOAN: a borrowing from French (
     morphologically adapted
                into Luxembourgish)
    - DE_LOAN: a borrowing from German (
     morphologically adapted
                into Luxembourgish)
    - EN_LOAN: a borrowing from English (
     morphologically adapted
                into Luxembourgish)
  Respond with ONLY the label on the first line and
      a one-sentence
   justification on the second line.

  Example:
    Sentence: D'**Pompjeeën** hunn de Brand
     schnell ënnert
              Kontroll bruecht.
    Assistant: FR_LOAN
    Justification: 'Pompjeeën' derives from French
     'pompier',
      adapted with the Luxembourgish plural suffix
     "-en" and
      spelling "ee" for /e:/.

User (neology task):
  Given the following Luxembourgish sentence and
     the highlighted
  word, decide whether this token should be
     treated as a lexical
  innovation in Luxembourgish.
  Answer YES or NO, followed by one sentence of
     explanation.

  Sentence: [Luxembourgish sentence containing **
     TOKEN**].
```

Knowledge-augmented prompts add morphological information derived from LuxBorrow and LOD. In the `KG-flat` conditions, the user message begins with a preamble *"According to the LOD, the following morphological adaptation patterns are productive in Luxembourgish:"*, followed by up to twenty globally fixed pattern entries. Each entry lists a pattern name, its type (morphological, orthographic, or lexical), the donor language, and up to three example pairs, for example *"éiere → er, type. morph, donor. FR, e.g. abordéieren ← aborder"*. This global pattern block is identical in `KG-flat` and is appended to every instance, which contrasts with `KG-graph`, where the context is an instance-specific LKG subgraph as described in Section 3.2.

To quantify the contribution of individual LKG components, we define six ablation variants that selectively remove lexicon attestation, etymology chains, synonym links, analogical examples, or contrastive patterns, as well as a `lex-only` condition that keeps only dictionary-style information without graph structure. Together, the eleven base strategies and six ablations define 17 prompt setups per task. Applied to both borrowing classification and neology detection, this yields 34 task-specific evaluation settings per model.

3.4. Models

All experiments are conducted with instruction-tuned, general-purpose LLMs accessed through an OpenAI-compatible endpoint (OpenRouter API). We deliberately treat the models as frozen black boxes and rely exclusively on prompting; no fine-tuning is performed.

We consider three model sizes: Gemma 3 12B (`google/gemma-3-12b-it`) as a small model, Gemma 3 27B (`google/gemma-3-27b-it`) as a medium model, and Llama 3.3 70B Instruct (`meta-llama/llama-3.3-70b-instruct`) as a large model. All runs use a temperature of 0.0 and a maximum output length of 1,024 tokens to enforce deterministic, label-complete responses. Combining three models with 34 prompt configurations yields 102 evaluation settings, each applied to the full LexNeo-Bench of 3 050 instances.

3.5. Evaluation protocol

Model outputs often contain explanations or formatting artifacts, so we post-process responses to recover a single canonical label per instance. We strip explicit reasoning blocks (for example between `<think>` and `</think>`), then examine the first and last non-empty lines and map them to one of the allowed labels using a small normalization dictionary (for example, `FRENCH`, `FR`, or `FR_loanword` all map to `FR_LOAN`, while `LUXEMBOURGISH` or `LB` map to `NATIVE`). Outputs that cannot be unambiguously resolved are marked as `PARSE_ERROR` and omitted from metric computation; we report their frequency separately.

For the borrowing classification task, we report accuracy, balanced accuracy, macro- and weighted-F1 over the active classes, as well as per-class precision, recall, F1, and confusion matrices. In addition, we analyze two derived sub-tasks: a binary native versus borrowed decision (collapsing `FR_LOAN` and `DE_LOAN`) and donor-only discrimination between `FR_LOAN` and `DE_LOAN`. For the neology task, we treat `YES` as the positive label and report accuracy, precision, recall, and F1, with additional breakdowns by donor language. The gold label is derived directly from the primary LuxBorrow borrowing annotation: tokens annotated as `FR_LOAN`, `DE_LOAN`, or `EN_LOAN` are mapped to `YES` (lexical innovation), and `NATIVE` tokens to `NO` (see Section 3.1). All metrics are computed on the same fixed test set.

Temporal robustness is assessed by comparing accuracies on established versus recent items and reporting the absolute gap. Finally, the evaluation pipeline supports resumable execution. Prediction files are incrementally extended when experiments are restarted, which makes large grids of runs robust to interruptions without recomputation.

4. Results

4.1. RQ1. Borrowing classification performance

Table 1 summarizes three-way borrowing classification accuracy and macro F1 across models and prompt strategies. Without a structured linguistic context, performance remains modest. In the zero-shot baseline, accuracy ranges from 24.5% for Gemma 3 12B to 34.7% for Llama 3.3 70B, and more elaborate non-KG prompts, such as Few-shot, remain below 42% across all models.

Since models choose from four output labels, a random baseline yields 25% accuracy; zero-shot performance ranges from 24.5% to 34.7%, indicating that parametric knowledge alone barely exceeds chance.

Introducing a structured linguistic context via the KG-graph condition changes this picture sharply. With KG-graph, accuracy rises to 81.0% for Gemma 3 12B, 71.4% for Gemma 3 27B, and 71.3% for Llama 3.3 70B, and macro F1 exceeds 0.55 for all models, peaking at 0.634 for Gemma 3 12B. The simpler KG-flat variant, which exposes only a global list of morphological patterns, does not close this gap and behaves similarly to non-KG baselines. The improvement, therefore, stems from instance-specific retrieval rather than merely reminding the model that borrowing patterns exist. Taken together, these results answer RQ1 by showing that structured, token-level linguistic context is necessary to achieve robust borrowing classification in Luxembourgish.

Table 1: Acc. and macro F1 (in %), and KG gain Δ_{KG} (percentage points), defined as the accuracy difference between KG-graph and zero-shot.

Model	Prompt	Acc.(%)	Macro F1	Δ_{KG}
Gemma 3 12B	Zero-shot	24.5	22.3	
	Few-shot	38.3	30.3	
	KG-flat	30.3	26.2	
	KG-graph	81.0	63.4	+56.5
Gemma 3 27B	Zero-shot	31.4	19.7	
	Few-shot	38.0	29.8	
	KG-flat	33.7	22.6	
	KG-graph	71.4	55.9	+40.1
Llama 3.3 70B	Zero-shot	34.7	27.9	
	Few-shot	38.0	29.0	
	KG-flat	36.3	27.9	
	KG-graph	71.3	55.7	+36.6

4.2. RQ2. Per class performance and donor bias

To understand where the gains from KG-graph conditioning arise, Figure 2(A) reports per class F1 under the KG-graph prompt. All three models achieve strong F1 scores for French and German borrowings. Gemma 3 12B reaches 0.920 for FR_LOAN (French borrowing) and 0.840 for DE_LOAN (German borrowing); Gemma 3 27B reaches 0.880 and 0.750 respectively, and Llama 3.3 70B scores 0.921 and 0.791. Performance on NATIVE items is more variable. Gemma 3 12B maintains a solid 0.777 F1, whereas Llama 3.3 70B drops to 0.515, suggesting that the larger model overfits to donor cues and sometimes over-predicts borrowing for genuinely native words.

Confusion patterns show a marked donor asymmetry. In Figure 2(b), the dominant error is French-origin items misclassified as German-origin (FR_LOAN→DE_LOAN: 18,142 cases across all model and prompt combinations), which is 4.6× more frequent than the reverse direction (DE_LOAN→FR_LOAN: 3,946). Native Luxembourgish items are also misattributed to German borrowings (20,662) substantially more often than to French borrowings (5,944), indicating an overall tendency to overpredict DE_LOAN. This pattern is consistent with potential lexical/orthographic overlap between French- and German-origin forms in Luxembourgish, although other factors (e.g., class priors or KG coverage) may also contribute. Overall, these results support RQ2: while KG-graph improves borrowing recognition, donor identification remains skewed toward German across model and prompt settings.

EN_LOAN as a distractor label. Although EN_LOAN is absent from the evaluation set (only 24 source instances, below the 50-instance threshold), we retain it as a valid output label to probe whether models project English-origin hypotheses onto tokens that are in fact native or borrowed from French or German. Table 2 reports how often each model predicts EN_LOAN and which true class absorbs those false positives.

Table 2: EN_LOAN false-positive analysis. **EN_{pred}** is the total number of EN_LOAN predictions; columns show the true-class breakdown of those predictions. **Rate** is the proportion of all valid predictions assigned to EN_LOAN.

Model	Prompt	EN_{pred}	→NAT	→FR	→DE	Rate
Gemma 12B	Zero-shot	1 001	392	196	387	32.8%
	Few-shot	238	103	29	97	7.8%
	KG-flat	627	254	80	275	20.6%
	KG-graph	128	107	5	12	4.2%
Gemma 27B	Zero-shot	312	130	47	129	10.2%
	Few-shot	212	96	23	89	7.0%
	KG-flat	299	125	31	136	9.8%
	KG-graph	167	128	6	28	5.5%
Llama 70B	Zero-shot	363	196	37	125	12.0%
	Few-shot	91	38	6	43	3.0%
	KG-flat	87	32	7	45	2.9%
	KG-graph	264	222	7	32	8.7%

Two patterns stand out. First, without structured context, models frequently over-predict EN_LOAN: Gemma 12B assigns it to nearly a third of all

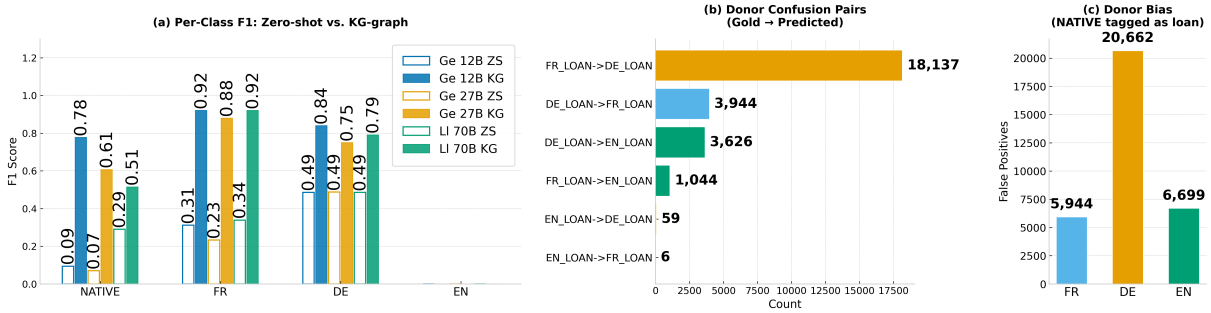

Figure 2: (a) Per-class F1 for zero-shot and KG-graph by model (GE12, GE27, & LL70 denote Gemma 3 12B, Gemma 3 27B, & Llama 3.3 70B, respectively.). (b) Top donor confusion pairs. (c) False-positive rates on NATIVE: proportion of NATIVE tokens predicted as loans.

tokens under zero-shot prompting. KG-graph prompting reduces the EN_LOAN rate by 78–87% for the Gemma models (from 32.8% to 4.2% for Gemma 12B, and from 10.2% to 5.5% for Gemma 27B), confirming that structured linguistic context suppresses spurious English-origin hypotheses. Second, across all models and prompt conditions, the majority of false EN_LOAN predictions fall on genuinely NATIVE tokens rather than on French or German borrowings. Under KG-graph, this concentration intensifies: 84% of Gemma 12B's and 77% of Gemma 27B's residual EN_LOAN predictions fall on NATIVE tokens. This suggests that when models lack donor-specific evidence, they default to an English-origin hypothesis for unfamiliar Luxembourgish words, a bias consistent with English's dominance in multilingual pretraining corpora.

Retaining EN_LOAN as a distractor label therefore serves a diagnostic purpose: it exposes this bias and provides a measurable signal of how effectively structured context can counteract it.

4.3. RQ3. Ablating KG components

Figure 3 shows the effect of removing individual components from the KG-graph prompt. The full KG-graph condition reaches 81.0%, 71.4%, and 71.3% accuracy for the three models (Gemma 3 12B, Gemma 3 27B, and Llama 3.3 70B). Removing etymological information (No Etymology) reduces accuracy to 78.9% for Gemma 3 12B, 58.4% for Gemma 3 27B, and 69.2% for Llama 3.3 70B. Dropping analogical examples (No Analogues) has a similarly strong impact, especially on the 27B model, where accuracy decreases by roughly 13 percentage points.

By contrast, removing synonym links or contrastive patterns changes performance only marginally, within ± 0.3 points of the full KG-graph condition. A Lexicon-only variant that keeps dictionary entries but discards graph structure clearly outperforms non-KG baselines, yet remains 6–19 points behind the full graph, which suggests that donor chains and pattern-sharing analogues carry most of the useful signal, while long definitions may introduce noise. In some settings, accuracy even improves slightly when the lexicon text is removed, but the graph structure is kept, reinforcing that relational structure is more valuable than raw definitional prose.

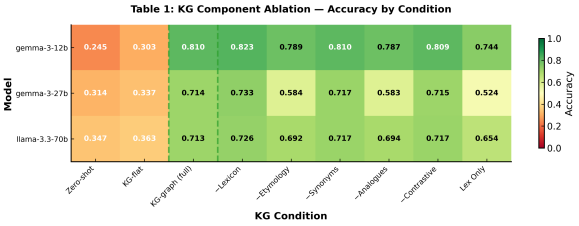

Figure 3: Impact of KG components on borrowing accuracy by model and KG-graph ablation condition.

4.4. Model scale and benefit from KG

Table 1 compares zero-shot and KG-graph accuracy by model size. Zero-shot accuracy grows modestly with scale, from 24.5% for Gemma 3 12B to 34.7% for Llama 3.3 70B, but under KG-graph the ranking inverts: Gemma 3 12B reaches 81.0%, while Gemma 3 27B and Llama 3.3 70B plateau at 71.4% and 71.3%, respectively. The KG gain Δ_{KG}, defined as the accuracy difference between KG-graph and zero-shot, decreases monotonically with model size: +56.5, +40.1, and +36.6 percentage points. A supplementary log-scale visualization of this trend is provided in Appendix 4.

The reversal is KG-specific. Gemma 12B does *not* generally outperform Gemma 27B. Under zero-shot (31.4% vs. 24.5%), few-shot (38.0% vs. 38.3%, essentially tied), and KG-flat (33.7% vs. 30.3%), Gemma 27B matches or exceeds

Gemma 12B. The reversal occurs exclusively under KG-graph (+9.5 pp in favor of 12B), ruling out a general quality advantage of the smaller model and localizing the effect to how each model utilizes instance-specific structured context.

Mechanism: NATIVE over-prediction by larger models. Under KG-graph, all models achieve high recall on FR_LOAN ($\geq$0.860) and DE_LOAN ($\geq$0.961), but NATIVE recall drops sharply with scale: 0.639 (12B), 0.447 (27B), 0.349 (70B). Precision on NATIVE remains above 0.94 for all models, so larger models predict NATIVE correctly when they do—but they predict it far too rarely, over-attributing borrowing status to native words. KG ablations confirm this asymmetry: removing etymology or analogues costs Gemma 27B $\sim$13 pp but Gemma 12B only $\sim$2 pp, showing that the larger model falls back on parametric borrowing priors when graph evidence is incomplete.

This pattern is consistent with findings on parametric–contextual knowledge conflicts (Longpre et al., 2021; Xie et al.): larger models develop stronger internal representations of French and German items during pre-training, which compete with KG-supplied evidence and lead to over-attribution of donor origins. The smaller model, lacking such entrenched priors, defers more faithfully to the structured context.

In this analysis, we use publication dates as a diachronic proxy to contrast more established items with more recent adaptations (pre-2015 vs. post-2015). The graph encodes structural origin information (donor language, morphological patterns, analogues) but not explicit recency cues such as frequency trajectories or first-attestation dates. Under this temporal split (see Supplementary), KG-graph is the most temporally robust condition, with recent vs. established gaps of only 0.7–2.8 pp.

Under the KG-graph condition, Gemma 3 12B achieves 81.4% accuracy for established items and 80.7% for recent ones; Gemma 3 27B achieves 73.1% and 70.3%; Llama 3.3 70B reaches 72.4% and 70.6%. These results indicate that the graph captures structural regularities that transfer to more recent lexical items, even if such items are under-represented or missing in the models' pre-training data. Among all prompt strategies, KG-graph is the least affected by recency, suggesting that structured linguistic context can partially compensate for gaps in parametric training data; a full breakdown by model and prompt is provided in the supplementary material.

4.5. Neology detection

The neology decision task, which collapses all borrowings into a single lexical-innovation class versus native items, behaves very differently from borrowing classification. Table 3 reports accuracy and $F1_{\text{neo}}$ for the "neologism" class by model and prompt strategy. Here, few-shot prompting is consistently the most effective strategy. Gemma 3 12B reaches 48.5% accuracy and $F1_{\text{neo}} = 0.509$, Gemma 3 27B reaches 49.2% and 0.524, and Llama 3.3 70B achieves 40.8% and 0.254. In contrast, the KG-graph condition substantially degrades performance. Accuracy falls to 34.2%, 30.3%, and 30.5% for the three models, and $F1_{\text{neo}}$ for Llama 3.3 70B drops close to zero.

This divergence is in line with how the linguistic knowledge graph is constructed. The graph encodes origin and structural information (donor language, morphological pattern, analogues, native synonyms), which are exactly the cues needed for borrowing classification, but largely orthogonal to *recency*. Deciding whether a word counts as a lexical innovation requires diachronic evidence, such as frequency trajectories, first attestation dates, or domain-specific usage shifts, none of which are currently exposed in the graph. As a result, the additional context encourages models to reason about *where* a word comes from rather than *when* it entered the language, which can mislead them in borderline cases.

Table 3: Neology decision performance by model and prompt. Accuracy and $F1_{\text{neo}}$ for the "neologism" class.

Model	Prompt	Acc. (%)	$F1_{\text{neo}}$
Gemma 3 12B	Zero-shot	41.2	0.308
	Few-shot	48.5	0.509
	KG-graph	34.2	0.042
Gemma 3 27B	Zero-shot	45.6	0.429
	Few-shot	49.2	0.524
	KG-graph	30.3	0.064
Llama 3.3 70B	Zero-shot	36.7	0.127
	Few-shot	40.8	0.254
	KG-graph	30.5	0.012

4.6. Binary native versus borrowed

Finally, we collapse the four-class label space into a binary decision and ask models to distinguish native Luxembourgish words from any type of borrowing. Under the KG-graph condition, all models reach high performance. Gemma 3 12B attains 85.5% accuracy and $F1 = 0.902$, Gemma 3 27B reaches 78.9% and 0.862, and Llama 3.3 70B reaches 75.5% and 0.845.

The contrast between the binary decision and the donor-specific four-way task suggests that the main residual difficulty lies in separating French from German borrowings, rather than in detecting

whether a token is lexically integrated at all. In other words, once the knowledge graph is available, knowing that a word is a borrowing is comparatively easy, while pinpointing the correct donor in a dense Luxembourgish, French, and German contact zone remains challenging. Detailed binary results for all prompt strategies are reported in the supplementary material.

5. Discussion

Our results show that off-the-shelf multilingual LLMs have limited awareness of how a small contact language integrates lexical borrowings, even when trained on large multilingual corpora. With four possible output labels, a random baseline yields 25% accuracy; zero-shot performance ranges from 24.5% to 34.7%, indicating that parametric knowledge alone barely exceeds chance. This observation aligns with work in contact linguistics and neology that emphasizes community entrenchment, dictionary listedness, and usage patterns over purely formal cues (Treffers-Daller, 2025; Chesley and Baayen, 2010; Wolfer and Klosa-Kückelhaus, 2023). The models do not spontaneously replicate the "Simple View" of borrowing as operationalized in lexicographic and corpus studies.

LexNeo-Bench complements earlier borrowing and anglicism corpora in Spanish and other languages (Alvarez-Mellado, 2020, 2021; Mellado et al., 2021; Álvarez-Mellado et al., 2025; Álvarez Mellado, 2020; Kevers, 2022) by exposing LLMs to a dense Luxembourgish, French, and German contact zone where orthographic and morphological integration is pervasive (Adda-Decker et al., 2008; Lavergne et al., 2014; Anastasiou, 2022). The strong gains from structured knowledge-graph prompting suggest that models can make fine-grained borrowing decisions once they are supplied with token-specific morphological patterns, donor labels, and analogical examples. This mirrors gains observed when injecting gazetteers and knowledge bases into NER and entity-centric tasks (Tan et al., 2023; Chen et al., 2022) and supports the view that community lexical resources remain crucial even in the LLM era (Tomaszewska et al., 2025; Hosseini-Kivanani, 2025).

At the same time, our neology decision results highlight that structural donor information alone does not solve diachronic questions. LLMs perform best with few-shot prompting that clarifies the task mapping (borrowings count as lexical innovations), while knowledge-graph prompts, which were designed for borrowing classification, can even harm performance. This gap reflects broader findings on LLM-based neology detection and "LLM neologisms" that arise from tokenization and encoding artifacts rather than organic community usage (Iwamoto and Kanayama, 2024; Zheng et al., 2024). For small languages with sparse written production and heavy code mixing (Plum et al., 2024; Adda-Decker et al., 2008), separating genuine innovations from long-standing borrowings remains challenging without explicit temporal signals or external diachronic corpora.

Our study has several limitations: First, LexNeo-Bench is derived from a single edited news source, so it under-represents informal registers and spoken discourse. Second, borrowing labels rely on an automatic pattern pipeline and dictionary signals, which may misclassify borderline items or miss emerging forms in under-documented domains. Third, we evaluate only three instruction-tuned models with frozen prompts, so conclusions about model scale and architecture should be treated as tentative. Finally, the benchmark focuses on token-level decisions and does not directly measure how LLMs handle borrowing in generation, for example, in spelling correction or style transfer. Addressing these limitations will require extending the benchmark to other genres, adding human validation for difficult cases, and coupling classification with controlled generation tasks.

6. Conclusion and Future Work

We introduced LexNeo-Bench, a token-level benchmark derived from a borrowing-annotated Luxembourgish news corpus to probe how multilingual LLMs treat morphologically adapted borrowings. Across three models and 34 prompt configurations, zero-shot parametric knowledge stays near chance, whereas instance-specific linguistic knowledge graphs raise borrowing classification accuracy to about 71–81% and substantially improve binary native versus borrowed decisions, with the largest gains for the smallest model. This shows that structured lexical context can partly compensate for sparse pretraining in low-resource contact languages, while neology decisions remain difficult and are best supported by few-shot prompting in our experiments because recency is not encoded in the current graph. Future work will add explicit diachronic signals, extend LexNeo-Bench beyond edited news and Luxembourgish, and link token-level evaluation to downstream writing assistance to quantify the user-facing impact of borrowing misclassifications.

7. Acknowledgements

We thank RTL Luxembourg and Tom Weber for providing access to the news archive and for supporting its use for research purposes. This work

highly benefited from the collaborative network fostered by the **ENEOLI COST Action (CA22126)**, supported by COST (European Cooperation in Science and Technology), and also within the project LuxVoice (project reference 19205922) from the FNR.

8. Ethical and legal aspects

Data provenance and legal basis. The underlying corpus consists of online news articles published between 1999 and 2025 by a major Luxembourgish media outlet (RTL). The data were obtained under a formal research collaboration and processed under the outlet's terms of use and the applicable EU text and data mining provisions for non-commercial scientific research. No user accounts were accessed, no technical protection measures were circumvented, and we did not perform large-scale scraping of the public-facing website.

Data and code availability. Full prompt templates for all strategies and tasks, including the complete five-example few-shot prompt and the neology template, will be provided in the public GitHub: `github.com/NinaKivanani/LexNeo-Bench`.

Intellectual property and data release. All source articles remain under the copyright and database rights of RTL. Our preprocessing, annotation, and analysis operate on copies stored on secure institutional infrastructure; we do not redistribute the full text of the corpus. Instead, we release only derived artifacts that are not substitutable for the original content, including the annotation schema and pattern inventory, scripts to reproduce the pipeline on any legally obtained Luxembourgish news corpus, aggregate statistics and plots, and small illustrative excerpts.

Privacy and data protection. News articles naturally contain references to identifiable individuals. These mentions appear in material already made lawfully available online in the exercise of journalistic freedom, yet they still qualify as personal data. We do not link the corpus to external records, attempt to profile individuals, or infer sensitive attributes. All analyses are conducted at the token, sentence, or aggregate document level rather than at the level of specific persons. Data are stored and processed on secure servers, including national high-performance computing resources, in compliance with the GDPR and relevant national data protection requirements.

Intended use and potential impact. LexNeo-Bench reflects the editorial practices and topic mix of a single news provider and should not be treated as representative of all Luxembourgish language use. The benchmark and LKG are intended as descriptive tools for studying contact phenomena and for stress-testing multilingual LLMs on borrowing and neology, not as prescriptive standards for "correct" Luxembourgish. We explicitly discourage using these resources to police lexical borrowing, to stigmatize code-switching in everyday communication, or to draw strong sociolinguistic conclusions about specific groups. Any correlations between borrowing patterns and social or regional factors must be interpreted with caution to avoid reinforcing stereotypes or over-generalising from a single, institutionally edited source.

9. Bibliographical References

Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda. 2008. Developments of" lëtzebuergesch" resources for automatic speech processing and linguistic studies. In *LREC*.

Elena Alvarez-Mellado. 2020. An annotated corpus of emerging anglicisms in spanish newspaper headlines. In *Proceedings of the 4th workshop on computational approaches to code switching*, pages 1–8.

Elena Álvarez Mellado. 2020. *Lázaro: An extractor of emergent anglicisms in Spanish newswire*. Ph.D. thesis.

Elena Alvarez-Mellado. 2021. Extracting english lexical borrowings from spanish newswire. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 384–386.

Elena Álvarez-Mellado, Jordi Porta-Zamorano, Constantine Lignos, and Julio Gonzalo. 2025. Overview of adobo at iberlef 2025: Automatic detection of anglicisms in spanish. *Procesamiento del Lenguaje Natural*, 75:373–383.

Dimitra Anastasiou. 2022. Deliverable d1. 24 report on the luxembourgish language.

Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022. Ustc-nelslip at semeval-2022 task 11: Gazetteer-adapted integration network for multilingual complex named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1613–1622.

Paula Chesley and R Harald Baayen. 2010. Predicting new words from newer words: Lexical borrowings in french. *Linguistics*, 48(6).

Nina Hosseini-Kivanani. 2025. A hybrid framework for neologism validation using llms and lexical knowledge graphs. In *1st International Workshop on Terminological Neologism Management, NeoTerm*, pages 1613–0073.

Nina Hosseini-Kivanani and Fred Philippy. 2026. Luxborrow: From pompier to pompjee, tracing borrowing in luxembourgish. *arXiv preprint arXiv:2603.10789*.

Ran Iwamoto and Hiroshi Kanayama. 2024. Llm neologism: Emergence of mutated characters due to byte encoding. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 24–29.

Laurent Kevers. 2022. Coswid, a code switching identification method suitable for under-resourced languages. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 112–121.

Thomas Lavergne, Gilles Adda, Martine Adda-Decker, and Lori Lamel. 2014. Automatic language identity tagging on word and sentence-level in multilingual text sources: a case-study on luxembourgish. In *LREC*, pages 3300–3304.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7052–7063.

Elena Álvarez Mellado, Luis Espinosa Anke, Julio Gonzalo Arroyo, Constatine Lignos, and Jordi Porta Zamorano. 2021. Overview of adobo 2021: Automatic detection of unassimilated borrowings in the spanish press. *Procesamiento del Lenguaje Natural*, 67:277–285.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. Luxbank: The first universal dependency treebank for luxembourgish. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 30–39.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, et al. 2023. Damonlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2014–2028.

Aleksandra Tomaszewska, Dariusz Czerski, Bartosz Żuk, and Maciej Ogrodniczuk. 2025. Neon: A tool for automated detection, linguistic and llm-driven analysis of neologisms in polish. In *International Conference on Computational Science*, pages 318–326. Springer.

Jeanine Treffers-Daller. 2025. The simple view of borrowing and code-switching. *International Journal of Bilingualism*, 29(2):347–370.

Sascha Wolfer and Annette Klosa-Kückelhaus. 2023. Tracking the acceptance of neologisms in german: Psycholinguistic factors and their correspondence with corpus-linguistic findings. *Humanities and Social Sciences Communications*, 10(1):1–10.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

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.

10. Language Resource References

Zenter fir d'Lëtzebuerger Sprooch. 2025. *Lëtzebuerger Online Dictionnaire (LOD)*. Official reference dictionary for Luxembourgish.

11. Appendices

11.1. Supplementary visualization of KG gain

KG gain is defined as the accuracy difference between KG-graph and zero-shot prompting. The gain decreases monotonically with scale, from +56.5 percentage points for Gemma 3 12B to +40.1 for Gemma 3 27B and +36.6 for Llama 3.3 70B.

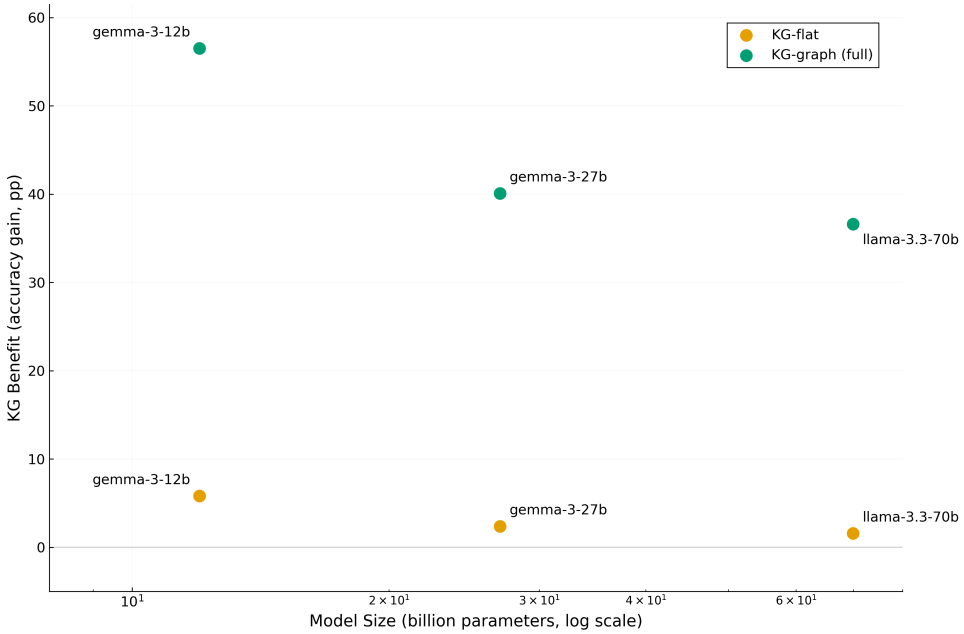

Figure 4: Supplementary visualization of KG gain Δ_{KG} by model size on a log-scaled x-axis.

Lexical Innovation in Business Colour Idioms: Evidence from Large Language Models in Five Languages

Giedre Valunaite Oleskevicienė[1], Ágnes Abuczki[2], Ganit Richter[3], Berat Ujkani[4], Vera Moitinho de Almeida[5], Pedro Madeira[6]

[1]Faculty of Human and Social Studies of Mykolas Romeris University; [2]Károli Gáspár University of the Reformed Church in Hungary; [3]Management Information Systems, School of Business Administration, The College of Management Academic Studies; [4]Faculty of Mechanical and Computer Engineering of University "Isa Boletini" in Mitrovica; [5]Faculty of Arts and Humanities of the University of Porto; [6] Independent Researcher

[1]Ateities 20, Vilnius, Lithuania; [2]Reviczky u. 4. 1088 Budapest, Hungary; [3]Elie Wiesel 2, Rishon LeTsiyon, Israel; [4]Ukshin Kovacica, 40000 Mitrovica, Kosovo; [5]Via Panorâmica, s/n, 4150-564 Porto, Portugal

[1]gvalunaite@mruni.eu, [2]abuczki.agnes@kre.hu, [3]ganitri@colman.ac.il, [4]berat.ujkani@umib.net, [5]vmoitinho@letras.up.pt, [6]pedromadeira1@gmail.com

Abstract

Lexical innovation refers to the process of creating new lexical items, enabling languages to adapt to evolving socio-cultural and material realities. The domains of business, economics, and finance are among the most productive ones of lexical innovation. The present research study lies at the intersection of lexical innovation, idiomaticity, and large language model (henceforth, LLM) research and investigates lexical productivity, semantic shift, and globalisation (Anglocentric changes) in business-related colour idioms by comparing human translation and annotation with the output of LLMs. The current experiment involves an initial study carried out for five languages: English (the pivotal one), Albanian (AL), Hebrew (HE), Hungarian (HU), Lithuanian (LT), and Standard European Portuguese (PT). The research results reveal that LLMs show high mutual agreement, but the agreement with humans is lower. The internal consistency of LLMs reflects shared Anglocentric metaphor encoding rather than convergence toward human idiomatic usage. It demonstrates that human expertise remains essential for high-quality idiomatic translation, particularly for culture-specific expressions.

Keywords: lexical innovation, LLMs, idioms

1. Related Research

Lexical innovation is a process of creation or adaptation of words or terms that allows languages to adapt to constantly changing technological and scientific contexts, as well as to the global influence of the multiplicity of cultural and social contexts. The lexical innovation process involves the introduction of new words and meanings into the lexicon of a language and promotes linguistic evolution, allowing languages to dynamically adapt to changes by introducing new concepts representing new technologies and integrating social and cultural shifts (Armstrong, 2016). Frequently, lexical innovation is related to the emerging new technological domains and developing technical domains that are characteristic of the emergence of new concepts, which demand precise designators. However, lexical innovations are also inherent in informal and literary language. Grieve et al. (2018) distinguish three main types of lexical innovations: formal neologisms, which comprise completely new words in the language; neo-semantic innovations, which make the process of new meanings assigned to existing words; and borrowings and calques, which are imported units from other languages or morphological translations.

The contexts like business, finances, and economics are considered among the most productive of lexical innovation, leading to the introduction of new technical words (Llopis and Sánchez-Lafuente, 2012). The authors observe that lexical innovation in the field is mostly related to the English language, which has an impactful influence as a lingua franca and also introduces the main lexical innovations in the different fields of science, technology, and the world of business and finance. It is related to the dynamism of the business area, which traditionally introduces new forms of investment and innovative financial mechanisms in the major world financial hubs, accumulating wealth and implementing cutting-edge ways to make money. Economic and capital diversification processes lead to the need for lexical innovation to define new business practices. In short, lexical innovation in the field of business comprises both the introduction of new terms and the adaptation of existing ones to innovative business and social contexts.

Neology detection studies are closely related to the domain of semantic change, which has recently been researched through computational approaches. Traditionally, neology detection involves corpus-based frequency analysis, lexicon comparison, and rule-based morphological processing (Kerremans & Prokić, 2018). Tahmasebi and Borin (2018) discuss

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 39–46
16 MAY 2026.

computational techniques to tackle lexical semantic change by providing the semantic change types from the computational perspective, including lexical replacement, named entity change, role changes, and temporal changes. The authors discuss distributional and embedding-based methods to identify semantic change by applying semantic vectors, topic distributions, and contextualized neural embeddings.

2. Research Experiment and Methodology

The current study involves an experiment carried out for five languages: English (the pivotal one), Albanian (AL), Hebrew (HE), Hungarian (HU), Lithuanian (LT), and Standard European Portuguese (PT) (following the Portuguese Language Orthographic Agreement of 1990 - CPLP, 1990), aligning idiomatic business and finance terms and multiword expressions related to black colour. Colour idioms were chosen to be the subject of the study because colour terms particularly enable fast lexical innovation as they are cognitively salient and culturally shared. Colour idioms in business often emerge from socio-economic, political, and technological change or regulation (Prusak and Valūnaitė-Oleškevičienė, 2024; Malyuga and Aleksandrova, 2020), and they undergo semantic shift or domain re-specialisation (Alousque, 2011)—e.g., the meaning of the idiom "black swan" shifted from “rare-event theory” to “business risk." For these reasons, they are excellent targets for testing the management of idiomatic lexical innovations by large language models (LLMs).

First, data was collected. English colour idioms with black were selected using term bases and dictionaries, such as the Financial Times Lexicon and the Investopedia dictionary, among others. (For further references, see section 9. Language Resource References). The English colour idioms were saved in a shared spreadsheet and were manually complemented with their (most commonly used) counterparts in five languages by native speakers (of Albanian, Hebrew, Hungarian, Lithuanian, and Standard European Portuguese). As a consequence, a multilingual language resource of business colour idioms was created, which provided the human baseline in our research.

In the subsequent stage of the experiment, LLM-based generative AI chatbots were prompted to give the equivalents of the selected English idioms in the five languages under scrutiny. Three advanced LLMs (Claude 4.5 Sonnet, Gemini Pro, and GPT-5.2 Auto[1]) were used in the experiment to investigate their effectiveness and accuracy in multilingual term search in the context of business and finance.

The following prompt was given to each LLM: “Provide the equivalents of the English idiomatic expressions in LT, HU, HE, AL, and PT, and provide each of their direct, literal translations to English in the given table. Provide one equivalent for each idiom in each language (and one literal translation to English). Provide it in a downloadable, editable Excel file.” The English idioms were attached in a table format (without equivalents in other languages). Model responses were elicited via API calls, using the default decoding parameters of the models. The files generated in reply included the English idiom (original), LT equivalent and its literal translation to English; HU equivalent and its literal translation to English; HE equivalent and its literal translation to English; AL equivalent and its literal translation to English; and PT equivalent and its literal translation to English. Afterwards, the resulting three files were merged into a single master spreadsheet, where the columns were grouped by language so that GPT, Gemini, and Claude outputs could be read side-by-side for every language. Ultimately, a heatmap of consensus was generated by Claude that visualises the level of agreement (full, partial, no agreement) between the three LLMs (GPT, Gemini, and Claude) for (the equivalents of) each idiom across all five languages.

In the closing, interpretive phase, the output of the models was compared to human annotation, and the comparative evaluation of the outputs of the three models was performed by the authors on the task of providing equivalents of business idioms in several languages. It poses a great challenge for evaluation that there is no universally accepted framework to evaluate large language models.

3. Research Questions

3.1. Do LLMs provide the same equivalents as native speakers?

3.2. How do LLMs absorb, generate, and disseminate new lexical items? Do LLMs tend to keep the English term or translate it? (loanwords vs. translations; globalisation vs. localisation)

3.3. Which LLM provides the most colloquial equivalents and prefers loanwords?

3.4. Which LLM tends to translate the terms?

3.5. Are recent innovative colour idioms (such as Black Friday) more globalised (i.e., more direct loans from English) than older, traditional colour

[1] Claude 4.5 Sonnet was released in September, 2025, and it is advertised as the strongest model for building complex agents, with a very high level of reasoning. Similarly, Gemini 3 Pro was released in November 2025, described by Google as its most intelligent AI model yet, featuring enhanced reasoning and coding capabilities. The third model under evaluation, GPT-5.2, was released in December, 2025, and brings adaptive reasoning into everyday use by solving harder work tasks more effectively and with more polish, particularly in spreadsheet formatting.

idioms that are rooted in universal concepts (such as black hole)?

3.6. Do LLMs reinforce Anglocentric colour metaphors and lead to cultural imperialism across languages?

The biases and uneven coverage by LLMs may privilege neologisms in high-resource languages while under-representing or distorting innovation in smaller linguistic communities (Gallegos et al. 2024).

3.7. Does any model invent a phrase (hallucination) that does not exist?

In fact, LLMs not only absorb neologisms from their training data but also have the capacity to generate novel lexical items, metaphors, or hybrid forms in response to prompts (Iwamoto and Kanayama, 2024). When LLMs produce unattested lexical items, the question is under what conditions can these be classified as errors, creative neologisms, or emergent lexical proposals?

3.8. Are the LLMs consistent in terms of grammatical accuracy (definiteness of nouns, definite articles, capitalization)? Which LLM is the most consistent and grammatically rigid?

4. Research Results

When reviewing the multilingual side-by-side data in the master spreadsheet, several interesting patterns were identified that highlight the different features of the models and may distinguish them in terms of rendering business colour idioms across languages.

4.1. Globalisation vs. localisation (Loanword vs. translation)

The idiom “Black Friday” is one of the clearest differentiators among the different models.

Both GPT and Gemini showed a tendency to keep "Black Friday" as a loanword - e.g., in Portuguese (Black Friday), which reflects current colour idiom usage where the English term is dominant.
In contrast, Claude suggested the translation of “Sexta-feira Negra” in Portuguese. While literally correct, it relates to a disastrous day and not to the big shopping discounts day after Thanksgiving Day, suggesting Claude might prioritise linguistic purity over cultural usage.
In Hebrew, GPT used the transliteration (בלאק פרייידי - "Black Friday"), while Gemini and Claude used the translation (יום שישי השחור - “The Black Friday”).
LLMs often opted for literal translations—e.g., Gemini provided very precise literal translations in Hebrew, such as translating “squeezing” for “blackmail”. Similarly, GPT often defaulted to repeating the English idiom in the Literal Translation column—e.g., translating the literal meaning of Black Friday just as "Black Friday."

4.2. Grammatical nuance (definiteness, capitalisation, and case)

Gemini consistently used the definite form for days in Albanian (e.g., “E Premtja e Zezë”—"The Black Friday"). However, GPT and Claude sometimes toggled between definite and indefinite (“e premtja” vs. “e premtë”), or varied capitalisation in Albanian (“E Premtja” vs. “Premte”) and similarly Lithuanian. The research reveals Gemini's notable consistency.
As for “black box," GPT correctly employed hyphenation in the Portuguese translation "caixa-preta," whereas Claude and Gemini produced a spaced compound (“caixa preta”). None of the three LLMs conformed to the orthographic conventions stipulated by the 1990 Portuguese Language Orthographic Agreement (CPLP, 1990), specifically the requirement that days of the week in Portuguese be written in lowercase.

4.3. Idiom interpretations

In Lithuanian, Gemini provided a literal translation for “in the black": “Dirbti pelningai” (literally: "Work profitably"). The same was observed in Hebrew, בפלוס (literally: "In plus"), as well as in Hungarian, “nyereséges” (literally: “profitable"), and Claude opted for “nyereségesen” (literally: “profitably").
Concerning Portuguese, all three models correctly identified the unique Portuguese idiom “no azul” ("in the blue"), which is a strong indicator of cultural awareness because this is a unique colour idiom distinct from the English "black." However, these results refer to Brazilian Portuguese and not to standard European Portuguese, where one would expect translations such as “com saldo positivo” or “com lucro” (literally, "with positive balance” or “with profit," respectively).

Moreover, we have found several examples of semantic drift (Hamilton and Jurafsky, 2016) and metaphor extension (Lakoff and Johnson, 1980) in colour idioms, namely:

- Black → illegality (black market) → systemic risk of negative consequences (black swan)
- Green → meaning linked to ecology → ESG (Environmental, Social, Governance) → finance → branding ethics
- Blue → meaning linked to sustainable ocean resources (blue economy) → ESG (Environmental, Social, Governance) → finance → branding ethics

4.4. Vocabulary choice

For the "Blackleg" in Lithuanian, LLMs provided “Streiklaužys” (literally: "Strike-breaker") and similarly in Hungarian - “Sztrájktörő” (literally: "strike-breaker").

There was a slight difference in Portuguese for "blackleg," which included a minor grammatical difference between LLMs:

- Gemini: fura-greve (singular: strike-breaker).

- Claude, GPT: fura-greves (plural: strike-breakers).

A similar situation was observed in Albanian:

- Gemini: Thyerës i grevës (Breaker of the strike).
- GPT: Thyes Grevash (breaker of strikes—plural, indefinite).
- Claude: Thyesës i grevës (Breaker of the strike).

Concerning "Black economy" vs. "Black market economy," most languages use terms equivalent to "Shadow economy" (e.g., “šešėlinė ekonomika" in Lithuanian) or "Parallel economy" (e.g., “economia paralela" in Portuguese) to distinguish the broader economic concept from the specific "black market" (“mercado negro" in Portuguese).

All three LLMs provided distinct but correct translations for "black market economy" in Portuguese: "economia de mercado negro," "economia informal," and "economia subterrânea" (literally, "black market economy," "informal economy," and "underground economy," respectively).

As to "Black-Scholes," all three LLMs provided the Brazilian Portuguese translation “modelo Black-Scholes" instead of the standard European Portuguese “modelo de Black-Scholes” (literally, "Black-Scholes model” and “model of Black-Scholes," respectively).

Regarding “black money," GPT provided “dinheiro negro," while Claude and Gemini provided “dinheiro sujo” (literally, "black money” and “dirty money”). Both idioms are correct in Portuguese, although some authors (e.g., Silva, 2009; Hortelão Lopes, 2015) distinguish between "dinheiro negro" (i.e., money applied in illegal activities) and "dinheiro sujo" (i.e., money acquired from illegal activities).

4.5. Visualisation of the output of LLMs

The generated agreement heatmap is basically a visual overview with colour coding:

- Green (3) = All 3 systems agree
- Amber (2) = Two systems agree
- Red (1) = All systems differ

Agreement by language is represented by vertical ranking. The breakdown presents the distribution of agreement levels for each language, with separate bars for equivalent vs. literal translations. It shows that HU and LT perform best, while AL and HE have the most disagreement. The reason for this might be that HU and LT have larger training datasets, enhancing reasoning capabilities than AL, which is a low-resource language in the context of LLM training.

Agreement by idiom is represented in a horizontal bar chart ranking all 19 idioms by overall agreement score. The highest agreement is visible in terms like "black box," "Black Friday," and "black hole" (universal, technical, and cultural terms).

Most disagreement is related to terms requiring cultural adaptation, like "in the black," "blackleg," and "blackmail." The agreement scores range from ~1.8 to ~3.0 (perfect agreement).
The visualisations clearly show that the three AI systems agree most on:

- Technical financial terms like "Black-Scholes," “Black Monday," and “Black Tuesday," among others.
- Universal concepts and more common terms, such as "black box" and "black hole," show strong consensus across almost all languages.

Concerning languages, Hungarian and Lithuanian expressions show the most agreement, with Hungarian demonstrating 73.7% equivalent agreement, 78.9% literal agreement, and Lithuanian demonstrating 68.4% agreement on both.

The visualisation reveals most disagreement on culturally specific idioms requiring local adaptation, such as “in the black," "blackleg," and "blackmail."

Languages with supposedly limited training data, such as Albanian and Hebrew, show less agreement; for Albanian, only 42.1% full agreement, and for Hebrew, 47.4% full agreement on equivalents. Albanian shows more red/yellow areas, indicating that the models struggle to agree on the definitive grammatical form (definite vs. indefinite articles, e.g., “e premtja” vs. “premtë”).

The models partially agree on several Portuguese idioms, where the heatmap shows yellow, as GPT and Gemini prefer the English loanword, e.g., "Black Friday," while Claude prefers the literal translation "Sexta-feira Negra."

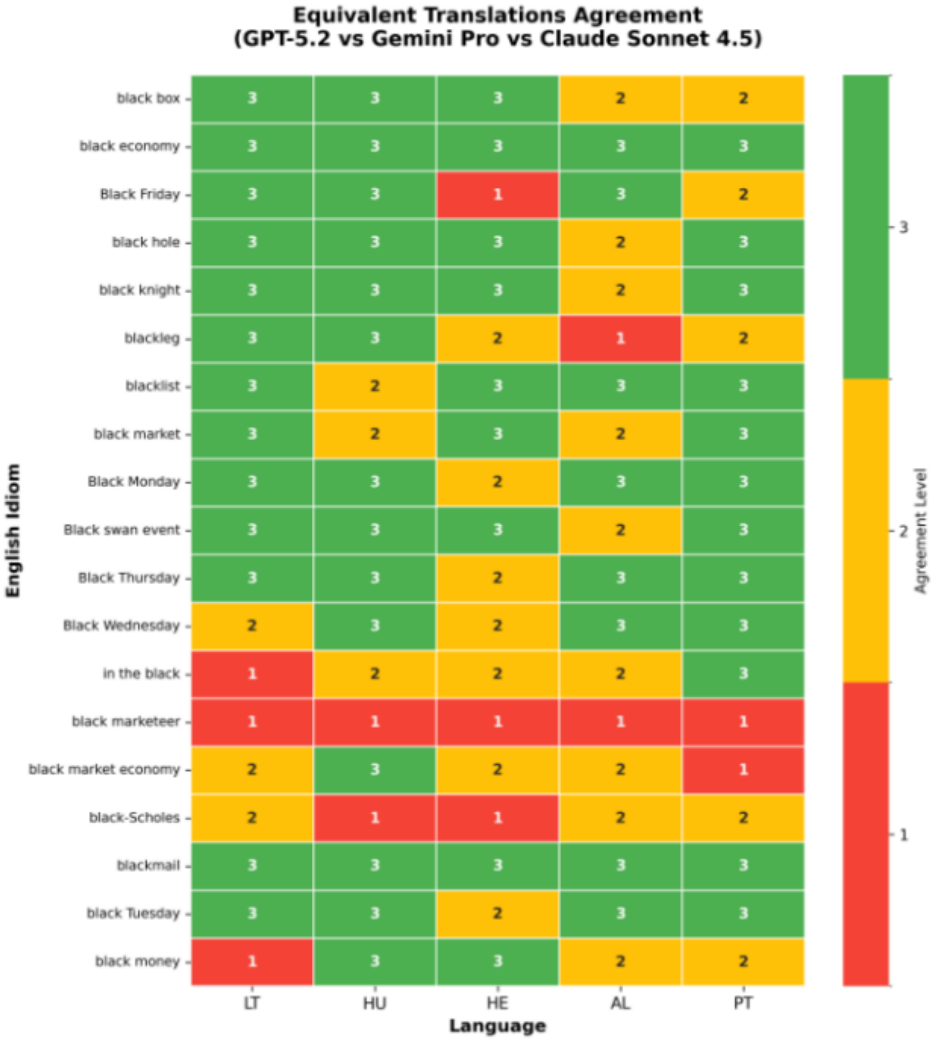

Figure 1: LLM consensus heatmap

4.5. Comparing Human Baseline with LLM Output

The results of the experiment demonstrate strong agreement among LLMs but less consistent agreement with human interpretation of colour idioms. The heatmap below compares the human-provided idioms against the collective output of the three LLMs (GPT, Gemini, Claude).

Green colour indicates high consensus, which indicates that all three LLMs produced the same translation as the human annotator. This indicates the idiom is well-established and standard in that language.

Yellow and orange slots indicate partial agreement, showing that some LLMs matched human choices, while others differed (e.g., using a loanword like "Black Friday" instead of the translated term).

The red color demonstrates disagreement where none of the LLMs matched the human translation. This frequently occurs when the human baseline states "No equivalent," but the LLM attempts to force a literal translation.

Human–LLM agreement clusters around partial agreement (yellow) rather than green.

The human choices demonstrate the use of nuanced local idioms (e.g., “no azul” in Brazilian Portuguese, but not in Standard European Portuguese), and the LLMs may fail to capture such cases, though in this specific case, the LLMs actually performed well.

High agreement is registered in cases of technical or globally standardised idioms like "black hole," "black market," and "blacklist." Stable idiomatic expressions guide high human and LLMs alignment.

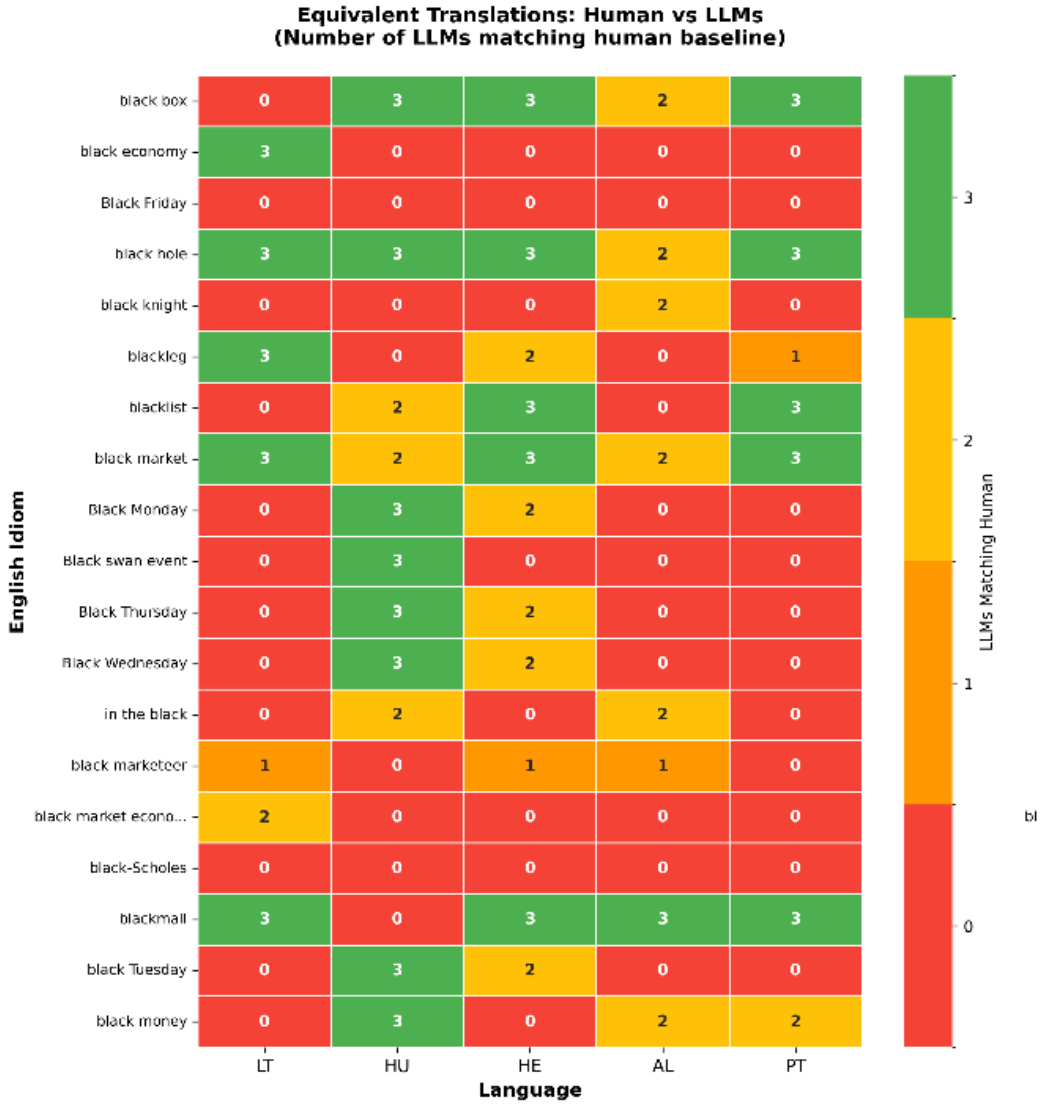

Figure 2: Human vs. LLM consensus heatmap

Human and LLMs' agreement breaks down in the case of idioms involving metaphor shifts, for example, in the case of *in the black, black economy, blackmail*, and *Black Friday*. In such cases, humans prefer functional or localised equivalents, while LLMs over-preserve English variants or loanwords.

Concerning languages in the dataset, Hungarian and Hebrew show higher human–LLM agreement. Portuguese and Lithuanian show more divergence, reflecting a stronger human preference for idiom substitution or localisation

High LLM–LLM consensus and less alignment of human-LLMs might mean that high cross-model agreement can coexist with systematic deviation from human idiomatic norms.

5. Conclusion

As a result of the comparative evaluation of the three models, it is claimed that, based on our small multilingual dataset, Gemini appears to be the most grammatically rigid (with a high level of consistency in terms of capitalisation and the use of definite articles). GPT seems to be the most colloquial out of the three models (preferring loanwords such as Black Friday in Hebrew and Portuguese). Claude leans towards academic translation (translating terms that might often be left in English). The visual heatmap shows that HU and LT perform best, while AL and HE have the most disagreement.

None of the three models invented a phrase that does not exist (no hallucinations).
These LLMs show higher mutual agreement than agreement with humans. In fact, LLMs are internally consistent, but that consistency reflects shared Anglocentric metaphor encoding rather than convergence toward

human idiomatic usage. The heatmaps clearly show that human expertise remains crucial for high-quality idiomatic translation, particularly for culture-specific expressions and less common language pairs.

6. Limitations

This is an ongoing research project first tested on a small dataset to carry out a pilot study, which inherently has a limitation of small-scale data. Another limitation is due to the deficient information on the technical specifications of the three examined models, because information such as model size is not publicly available. This presents a universal issue and an ongoing challenge when evaluating the performance of large language models.

7. Acknowledgments

This publication is based upon work from COST Action CA23147 GOBLIN—Global Network on Large-Scale, Cross-domain, and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

8. Bibliographical References

Alousque, N. I. (2011). A Semantic and Pragmatic Analysis of English Colour Idioms. *AFIAL/Journal of Semantics*, 20, 149–162.

Armstrong, J. (2016). The problem of lexical innovation. *Linguistics and Philosophy*, 39(2), 87–118.

CPLP (1990). *Acordo Ortográfico da Língua Portuguesa*. Comunidade dos Países de Língua Portuguesa (CPLP). www.cplp.org

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3), 1097–1179.

Grieve, J., Nini, A., & Guo, D. (2018). Mapping lexical innovation on American social media. *Journal of English Linguistics*, 46(4), 293–319.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (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) (pp. 1489–1501). Association for Computational Linguistics.

Iwamoto, R. & Kanayama, H. (2024). "Llm neologism: Emergence of mutated characters due to byte encoding". In *Proceedings of the 17th International Natural Language Generation Conference*, 24–29.

Kerremans, D., & Prokić, J. (2018). Mining the web for new words: Semi-automatic neologism identification with the NeoCrawler. Anglia, 136(2), 239-268.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Llopis, M. A. O., & Sánchez-Lafuente, Á. A. (2012). Deep into the discourse of the Spanish crisis. *Ibérica*, 23, 89–108.

Malyuga, E. N., & Aleksandrova, O. V. (2020). "Linguopragmatic Aspect of Idiomatic Expressions in English Business Discourse". In European Proceedings in Social and Behavioural Sciences.

Mateo, J. (2014). Neonyms for a crisis: Cognitive, terminological, and socio-pragmatic aspects in the translation of new financial terms into Spanish. *Babel*, 60(4), 405–424.

Prusak, B., & Valūnaitė-Oleškevičienė, G. (2024). Colour Idioms in Business Language. *Journal of Teaching English for Specific and Academic Purposes*, 12(3), 517–537.

Silva, G. M. (2009). "O crime de Branqueamento de Capitais e a Fraude Fiscal como crime pressuposto". In *Lavagem de dinheiro e injusto penal: Análise dogmática e doutrina comparada Luso-Brasileira*. Silva, L. N. & Bandeira, G. S. M. (Eds.). Curitiba, Juruá, pp. 239-253.

Tahmasebia, N., Borina, L., & Jatowtb, A. Survey of computational approaches to lexical semantic change detection. Computational approaches to semantic change, 1.

9. Language Resource References

Academia das Ciências de Lisboa (2023-). *Dicionário da Língua Portuguesa* [online]. Ana Salgado (ed.). Lisbon: Academia das Ciências de Lisboa/ILLLP. https://dicionario.acad-ciencias.pt/

ANACOM (2008). *Segundo relatório da CE sobre a aplicação da Directiva do acesso condicional* [News]. Autoridade Nacional de Comunicações (ANACOM). https://www.anacom.pt/render.jsp?contentId=680420

Anthropic. (2025). *Claude 4.5 Sonnet* [Large language model]. https://claude.ai/ (accessed on 30 December, 2025)

Berrance Simões, A. (1989). *Michaelis Dicionário Executivo: Administração, Economia, Marketing. Inglês-Português* (5th ed., Acordo Ortográfico). São Paulo: Comp. Melhoramentos.

Carvalho-Oliveira, J. M. & Fanha Martins, H. (2002). *A Vocabulary of Business, Accounting and Finance / Vocabulário*

Técnico Português-Inglês-Português. Lisbon: ISCAL - Instituto Superior de Contabilidade e Administração de Lisboa.

CGD (2022). *O Banco e Eu - Será que as suas finanças pessoais sobrevivem a um crash da bolsa? Saiba o que é e quais são as suas consequências*. Caixa Geral de Depósitos (CGD). https://www.cgd.pt/Site/Saldo-Positivo/o-banco-e-eu/Pages/crash-da-bolsa.aspx

FFMS (2024-). *O que é uma OPA hostil?*. Fundação Francisco Manuel dos Santos (FFMS). https://ffms.pt/pt-pt/direitos-e-deveres/o-que-e-uma-opa-hostil

Financial Times Lexicon https://markets.ft.com/glossary/searchTerm.asp?searchField=black&termId=

Fonseca, P. (2019). *A "terça-feira negra" que mudou o mundo há 90 anos: veja as imagens que ficaram para a História*. Visão - Mundo. https://visao.pt/atualidade/mundo/2019-10-29-a-terca-feira-negra-que-mudou-o-mundo-ha-90-anos-veja-as-imagens-que-ficaram-para-a-historia/

Google. (2025). *Gemini 3 Pro* [Large language model]. https://gemini.google.com/ and https://aistudio.google.com/ (accessed on 30 December, 2025)

Hortelão Lopes, E. A. (2015). *O ciclo vicioso do branqueamento de capitais: o caso português*. Bachelor thesis. Porto: Faculdade de Ciências Humanas e Sociais, Universidade Fernando Pessoa. http://hdl.handle.net/10284/4885

Investopedia dictionary (https://www.investopedia.com/financial-term-dictionary-4769738

Lietuvių kalbos institutas (2015-2026). *KALBA. Bendrinės lietuvių kalbos žodynas*. https://doi.org/10.35321/blkz

Mateus Ferreira (2018). *'Quarta-feira negra'. Como Soros se tornou no "homem que quebrou o Banco de Inglaterra"*. O Jornal Económico - Mercados. https://jornaleconomico.sapo.pt/noticias/quarta-feira-negra-como-soros-se-tornou-no-homem-que-quebrou-o-banco-de-inglaterra-355190/

Nunes Vicente, L. (2013). *Introdução à Matemática Financeira*. Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra. https://www.mat.uc.pt/~lnv/mf/mf.pdf

OpenAI, Inc. (n.d.). Models. *GPT-5.2.* [Large language model]. Platform.openai.com; OpenAI, Inc. https://platform.openai.com/docs/models/ (accessed on 29 December, 2025)

Pedro, C. (2009). *O "crash" que amolgou a economia do louco ano de 1929*. Jornal de Negócios - Economia. https://www.jornaldenegocios.pt/economia/detalhe/o_quotcrashquot_que_amolgou_a_economia_do_louco_ano_de_1929

Porto Editora (2009). *Dicionário de Inglês-Português* (5th ed., Acordo Ortográfico). Porto: Porto Editora.

Ricci, J. (1990). *Elsevier's Banking Dictionary - English/American, French, Italian, Spanish, Portuguese, Dutch, German* (3rd ed.). Amsterdam, New York: Elsevier.

Santos Silva, J. (n.a.). *Valores, cisnes e sabão (muito sabão)* [Press News]. Católica Lisbon School of Business & Economics. https://www.clsbe.lisboa.ucp.pt/noticias/valores-cisnes-e-sabao-muito-sabao

Where in Semantic Space Do Spanish Neologisms Emerge?

Bianca Delgado, Shira Wein
Amherst College
Amherst, MA, United States
{bdelgado27, swein}@amherst.edu

Abstract

English neologisms, or newly coined words, have previously been shown to emerge in sparser semantic neighborhoods (filling semantic gaps) and near other neologisms (in growing semantic areas). In this work, we investigate where in semantic space Spanish neologisms emerge, and whether this mirrors English neologism development. We find that Spanish neologisms, in comparison to non-neologisms, do indeed appear both nearer to other neologisms and further from non-neologisms. We additionally investigate the prevalence of loanwords from other languages through time in Spanish neologism production and manually assess the topics that appear as loanwords at four years: 1810, 1900, 1950, and 1990. Our findings show that on average, the Spanish neologisms in our dataset have fewer neighboring words in semantic space compared to non-neologisms and tend to cluster more tightly in the semantic space, indicating that patterns of neologism emergence span languages. This suggests that novel methods for neologism detection may be cross-lingually applicable, with these features serving as multilingual predictors of neologism emergence.

Keywords: vector semantics, Spanish, neologisms

1. Introduction

Neologisms are newly coined words that have been accepted in speech communities (Picone, 1996), and emerge across the world's languages. Ryskina et al. (2020) investigate how English neologisms relate to other words in semantic space, finding that English neologisms are more likely to (1) appear in sparser semantic neighborhoods (filling semantic gaps), and (2) emerge near other neologisms (in growing semantic areas). In this work, we investigate how typological diversity impacts how neologisms emerge, specifically examining whether these two findings by Ryskina et al. (2020) apply for Spanish neologisms. We perform this analysis through a temporal lens, performing this investigation on Spanish texts at various points in time and with multiple thresholds of "closeness" for semantic similarity (cosine similarities of at least 0.35, 0.45, and 0.55).

To address these research questions, we produce static embeddings of Spanish words appearing in the Google Ngram Viewer corpus (Michel et al., 2011), which contains frequencies of Spanish words used in books from 1500-2019. We classify these unigrams as neologisms or non-neologisms based on their proportion of usage before and after the year being analyzed. We then compare the embeddings of the Spanish words for each year via cosine similarity, counting how many words are at least as similar as the specified similarity lower bound. Then, we can use these counts to evaluate whether neologisms are more likely to emerge in (1) sparser areas, i.e. have fewer close semantic neighbors, and (2) growing semantic neighborhoods, i.e. have a higher proportion of neologisms as close semantic neighbors and tend to be grouped together in semantic space.

Next, we inspect clusters of Spanish neologisms to identify the topics that have grown in popularity over each of the various years, and assess the role that language contact has played in the creation of new words across time, by identifying the source language of the neologisms for each year.

We find that new Spanish words, like new English words, do indeed emerge in sparser and growing neighborhoods. We also find that language contact and globalization tend to impact the loanwords that appear in Spanish text over time. These findings indicate that neologisms tend to emerge in similar areas of semantic space across languages, given that these patterns appear to be consistent multilingually. This result opens up new avenues of multilingual neologism detection, which corresponds with one persistent challenge for large language models (LLMs) in the modeling of contemporary speech: unknown token handling.

2. Related Work

Neologisms reflect cultural, technological, and societal change. Consequently, it is important to be able to accurately detect and infer the meaning of neologisms from limited context in order to keep language models current and effective for real-world applications. In fact, Zheng et al. (2024) introduce NEO-Bench, a benchmark that serves to evaluate LLMs' robustness in handling neologisms, and conclude that LLMs are not yet fit to generalize on neologisms.

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 47–52
16 MAY 2026.

Prior work related to neologisms has largely focused on detecting their presence in corpora, in particular for English (Würschinger et al., 2016; McCrae, 2019; Zalmout et al., 2019). Kulkarni et al. (2018) propose leveraging the appearance of neologisms to help estimate when a document was written by tracking their appearance and frequency. In a non-English setting, prior work has investigated detecting neologisms in Persian (Megerdoomian and Hadjarian, 2010), Mandarin (Liu et al., 2013), French (Falk et al., 2014; Lejeune and Cartier, 2017), Russian (Lejeune and Cartier, 2017), and Japanese (Breen et al., 2018). Mizrahi et al. (2020) introduce a model which, rather than detecting or analyzing existing Hebrew neologisms, is designed to create new words with the goal of reducing reliance on loanwords.

In this work, we focus on characterizing the semantic qualities of Spanish neologisms, in particular for nouns. On the other hand, Rello and Basterrechea (2010) present the first system able to identify and conjugate Spanish verb neologisms, and Wein (2020) categorizes utterances in a Spanish language learner corpus as being neologisms, loanwords, or errors.

As discussed in Section 1, Ryskina et al. (2020) propose two hypotheses surrounding neologism emergence, in order to analyze neologisms through the lens of distributional semantics. Separate Word2Vec embeddings are trained on the Corpus of Historical American English (COHA; Davies, 2010) and the Corpus of Contemporary American English (COCA; Davies, 2008), and then aligned. Neologisms are identified as nouns that occur at least 20 times more frequently in the contemporary corpus, following Ryskina et al. (2020). Semantic density is measured by counting words that fall within certain cosine similarity thresholds ranging from 0.35 to 0.55, and frequency growth is calculated by averaging the change in frequency of a given word's neighbors over time. In concluding that both semantic sparsity and frequency growth serve as strong predictors, with frequency growth outperforming, this study offers valuable insight into neologism emergence for English words. These findings motivate our work on Spanish neologisms.

3. Methods

To test our hypotheses, we represent words using word embeddings and measure the semantic similarity between them. In doing so, we are able to define neighborhoods around each word via a similarity threshold in order to measure the density of the neighborhood as well as the presence of nearby neologisms.

3.1. Data

We utilize the unigram data from the 2020 Spanish-language Google Ngram Viewer corpus (Michel et al., 2011), which contains frequency counts of words in Spanish books published between 1500 and 2019. Each entry within this dataset contains a word and its frequency for each year.

Following Ryskina et al. (2020), before classifying neologisms, we filter the data to include only nouns using the part-of-speech tagger from the SpaCy package (Honnibal et al., 2020). Additionally, we filter out words beginning with capital letters as well as words with special characters or numbers. This preprocessing limits our dataset to just nouns, because they are an open-class part-of-speech, and helps reduce some of the noise from our large dataset by filtering out some named entities.

We identify neologisms at four cutoff years: 1810, 1900, 1950, and 1990. We select these four years given the amount of data available in each of those four intervals and the cultural shifts that occurred between those times. We then calculate the amount of times each word is used before and after each cutoff and their proportion of modern usage, which we define as the ratio of occurrences after the cutoff to occurrences before. We determine which words to label as neologisms based on their proportion of modern usage, ultimately labeling the 1,000 words with the highest proportion at each year as neologisms (we discuss examples of these neologisms in Section 4). We select the top 1,000, which follows Ryskina et al. (2020) and is affirmed by our qualitative assessment of the point at which the modern usage proportion drastically declines. The average ratios for these 1,000 highest-proportion items are as follows: 76,649 for 1810, 12,716 for 1900, 3,877 for 1950, and 564 for 1990. The decrease in these proportions over the years is an expected trend that likely reflects the amount of time each word had to accumulate usage after each cutoff year. For example, words that emerged around 1810 had over 200 years to become widely used, while words that emerged around 1990 had only a few decades until the end of the dataset in 2020.

3.2. Approach

In order to represent our words as vectors, we use static embeddings from fastText (Bojanowski et al., 2017),[1] as our dataset consists of isolated words and frequencies.

To address the first hypothesis, which is that neologisms tend to appear in less dense areas of the

[1] Specifically, we use the embeddings produced by the cc.es.300.bin Spanish-language model, which is pretrained on Common Crawl and Wikipedia data.

Language	Neologism	Non-neologism
Spanish	403	293429
English	198	95880
Portuguese	69	61794
Italian	61	52669
French	31	35193

Table 1: Source languages identified for neologisms and non-neologisms for words in the year 1990.

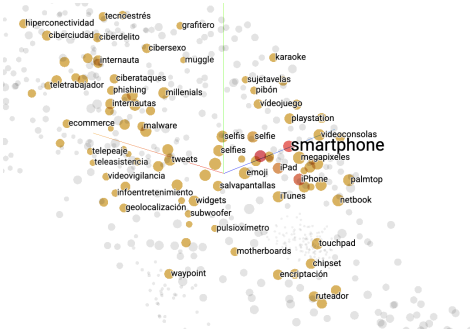

Figure 1: Technology Cluster: 1990 Cutoff. Words that are redder (darker) are closer in semantic space to smartphone.

semantic space, we calculate the cosine similarity between the each word and all other words in the dataset. Then, for each word, we count the number of similar words above a certain threshold. We choose to analyze several lower bound thresholds in order to explore varying degrees of similarity: 0.35, 0.45, and 0.55 following Ryskina et al. (2020); these thresholds are defined as the minimum cosine similarity that is required for a word to be considered "similar" to a given word, i.e. to be a part of that word's semantic neighborhood. This allows us to determine how many words are similar to the original word, and thereby ascertain the density of a given word's semantic space.

We operationalize the second hypothesis, which is that neologisms tend to arise in growing semantic neighborhoods, in terms of proximity of neologisms to each other. Accordingly, we perform the same calculation as for the first hypothesis, but now only focus on how many neologisms are near the word, indicating whether the area is growing.

Finally, to investigate the question of language contact's role in neologism creation, we perform language identification of the neologisms for each year. We leverage fastText's pretrained model for language identification, which supports 176 languages (Joulin et al., 2017).[2] We count the amount of words identified as belonging to each language by the model.

[2]Specifically, we use the lid.176.bin fastText model.

4. Results

First, we perform a qualitative analysis of the neologisms that emerge in the dataset at each year. We are able to notice clear trends in the types of words classified as neologisms. For the interval starting at 1810, we notice words like *fotografía* (photography), *vodka* (vodka), and *fútbol* (soccer/football) being labeled as neologisms. All of these words not only give insight into what terms are growing in popularity at the time, but are also examples of globalization's impact on language, as all of these terms are either loanwords or adapted from other languages—mostly from English, but with *vodka* being a Russian loanword. In the 1900-1950 data split, we see words like *telenovelas* (television soap opera), *ecologismo* (environmentalism), and *neurociencia* (neuroscience) reflecting a growing focus on scientific, environmental and media-related domains for the time period. For 1950-1990, we see *feminicidio* (femicide), *rockero* (rocker), *hippie* (hippie), *interculturalidad* (interculturality), and even *píxeles* (pixels), which reflects more modern developments. Finally, in the data split with cutoff year 1990 (which spans 1990 to 2020), we notice more contemporary concepts that are emblematic of 21st century advancements, such as *smartphone* (smartphone), *bitcoin* (bitcoin), and *ibuprofeno* (ibuprofen). Overall, this progression underscores the influence of globalization, technological innovation, and societal shifts on the Spanish language, revealing predictable patterns for neologism emergence. When visualizing the neologisms into cluster via T-SNE (Figure 1), we are able to observe similar clusters developing, including technologically- and scientifically-focused terms clustering together for each year.

When examining the amount of neologisms identified as belonging to each language for each cutoff year, we observe an increase in the proportion of neologisms that are non-Spanish words over time, with Portuguese and English appearing most frequently. In 1810, Portuguese is the most common non-Spanish source of neologisms, with English behind it. However, in 1900, English not only surpasses Portuguese, but continues to increase and ultimately makes up more than double the amount of Portuguese neologisms by 1990 (Table 1). This shift suggests a rise of English influence on Spanish vernacular.

Addressing the first hypothesis, our results (Table 2) show that on average, neologisms have fewer neighboring words in the semantic space compared to non-neologisms. Noticeably, the absolute counts are large, which is due both to the size of the dataset and to the use of a relatively low similarity threshold, such as 0.35, permit the inclusion of many loosely related words. For this

		Lower Bound		
Year	**Type**	**0.35**	**0.45**	**0.55**
1810	Neologism	497,530	73,111	10,147
1810	Non-neologism	693,953	569,252	324,618
1900	Neologism	473,241	104,741	17,723
1900	Non-neologism	693,985	569,210	324,608
1950	Neologism	461,496	139,012	28,781
1950	Non-neologism	694,001	569,165	324,593
1990	Neologism	488,120	233,502	102,150
1990	Non-neologism	693,965	569,039	324,495

Table 2: Average number of similar *words* within the threshold for neologisms and non-neologisms, at each lower bound. For example: for the year 1810, for all 1,000 neologisms the average number of similar words from the dataset in each neologism's neighborhood is 497,530 with a cosine similarity of $0.35 \leq x \leq 1$.

		Lower Bound		
Year	**Type**	**0.35**	**0.45**	**0.55**
1810	Neologism	37.4	9.25	2.59
1810	Non-neologism	3.44	0.309	0.045
1900	Neologism	34.0	8.53	2.53
1900	Non-neologism	7.19	1.18	0.190
1950	Neologism	32.7	8.60	2.55
1950	Non-neologism	15.8	2.87	0.336
1990	Neologism	57.4	21.6	6.00
1990	Non-neologism	53.6	13.3	1.87

Table 3: Average number of similar *neologisms* for neologisms and non-neologisms within each threshold. For example: for the year 1810, for all 1,000 neologisms the average number of similar neologisms in a given word's neighborhood is 37.376, for a cosine similarity of $0.35 \leq x \leq 1$.

reason, the values are best interpreted in comparison to one another rather than in isolation. The fact that neologisms have fewer neighboring words in the semantic space compared to non-neologisms indicates that neologisms are indeed more likely to emerge in sparser areas of the semantic space, suggesting that Spanish, like English, follows this supply-driven pattern. This is indicative of a cross-linguistic trend of conceptual gaps that exist in semantically sparse regions, which neologisms fill.

Conversely, when assessing the proximity of neologisms to each other (Table 3), the average similarity count is higher than that of non-neologisms across all years and thresholds. This suggests that new words tend to cluster more tightly in semantic space, in line with the second hypothesis, which outlines the demand-driven theory that neologisms emerge in areas of growing popularity. Our findings suggest that Spanish neologisms are also subject to such cultural trends and shifts, and thus that neologisms are not just filling gaps in the semantic space, but responding to increased demand in culturally relevant spaces.

These findings and the similarity of Spanish neologism emergence to English neologism emergence indicates that both semantic sparsity and growing popularity serves as a multilingual predictor of future neologism emergence.

5. Conclusion

In this work, we explore where in semantic space Spanish neologisms emerge in relation to other words. We specifically investigate whether Spanish neologisms are more likely to fill semantic gaps (thus appearing in sparser neighborhoods and having fewer close neighbors than non-neologisms do) and emerge in areas of growing popularity (thus being closer to other neologisms). As Ryskina et al. (2020) find for English, we find that Spanish neologisms do emerge in both sparser and growing semantic neighborhoods, suggesting that these phenomena carry across languages. Our qualitative analysis reveals that words related to technology and global politics regularly emerge as neologisms in our data. We additionally investigate the role of loanwords, finding that more English loanwords appear with increasing frequency over the centuries. Our findings motivate future work detecting multilingual neologisms given their relationships with other words and known neologisms. In particular, given that handling unknown tokens (such as neologisms) is a persistent challenge for LLMs, this work provides critical insight into how we may detect new words across languages, which would prove useful for enhancing performance of multilingual LLMs.

6. Limitations

We select the Michel et al. (2011) n-gram dataset because of its size, historical scope, and well-documented temporality metadata. The words appear as individual unigrams, and thus we are not able to leverage the context that the words appear in for our analysis (or use contextualized embeddings). A limitation of our approach is that semantic neighborhoods are computed in a modern embedding space. Because the dataset only contains isolated word usage per year and no text sequences, we cannot train embeddings that reflect the historical usage of each neologism. Future work using time-stamped corpora with contextual information and embeddings over time could address these issues, allowing analyses of historical semantic structure and the dynamics of emerging words, following the approach of Ryskina et al. (2020).

Further, our static embeddings do not allow us to determine whether semantically sparse regions existed prior to the emergence of neologisms or appear sparse because neologisms are newly introduced.

Additionally, we focus on Spanish nouns in particular, removing nouns with capital letters or special characters. While we filter out all nouns that begin with capital letters, some proper nouns remain in the dataset (such as “iPhone”).

Acknowledgments

We thank anonymous reviewers and members of the Amherst College NLP lab for their feedback. This work is supported by the Amherst College HPC, which is funded by NSF Award 2117377.

7. Bibliographical References

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.

James Breen, Timothy Baldwin, and Francis Bond. 2018. The company they keep: Extracting japanese neologisms using language patterns. In *Proceedings of the 9th Global Wordnet Conference*, page 163–171, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Mark Davies. 2008. The corpus of contemporary american english (COCA).

Mark Davies. 2010. The corpus of historical american english (COHA).

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in French newspapers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC‘14)*, pages 4337–4344, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Vivek Kulkarni, Yingtao Tian, Parth Dandiwala, 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.

Gaël Lejeune and Emmanuel Cartier. 2017. Character based pattern mining for neology detection. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 25–30, Copenhagen, Denmark. Association for Computational Linguistics.

Tsun-Jui Liu, Shu-Kai Hsieh, and Laurent Prevot. 2013. Observing features of PTT neologisms: A corpus-driven study with n-gram model. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 250–259, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

John Philip McCrae. 2019. Identification of adjective-noun neologisms using pretrained language models. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 135–141, Florence, Italy. Association for Computational Linguistics.

Karine Megerdoomian and Ali Hadjarian. 2010. Mining and classification of neologisms in Persian blogs. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 6–13,

Los Angeles, California. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The 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.

Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4918–4929, Online. Association for Computational Linguistics.

Michael Picone. 1996. *Anglicisms, Neologisms, and Dynamic French.* John Benjamins B.V.

Luz Rello and Eduardo Basterrechea. 2010. Automatic conjugation and identification of regular and irregular verb neologisms in Spanish. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 1–5, Los Angeles, California. Association for Computational Linguistics.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. 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.

Shira Wein. 2020. Classification and analysis of neologisms produced by learners of spanish: Effects of proficiency and task. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, page 88–91, Seattle, USA. Association for Computational Linguistics.

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.

Nasser Zalmout, Kapil Thadani, and Aasish Pappu. 2019. Unsupervised neologism normalization using embedding space mapping. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 425–430, Hong Kong, China. Association for Computational Linguistics.

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)*, page 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.

8. Language Resource References

Assessing the Pragmatic Competence of an LLM Regarding Novel Discourse Markers in Digital Communication

Ágnes Abuczki[1], **Giedre Valunaite Oleskevicienė**[2]

[1]Károli Gáspár University of the Reformed Church in Hungary; [2]Faculty of Human and Social Studies of Mykolas Romeris University
[1]Reviczky u. 4. 1088 Budapest, Hungary; [2]Ateities 20, Vilnius, Lithuania
[1]abuczki.agnes@kre.hu, [2]gvalunaite@mruni.eu

Abstract

The English language is changing faster than before, partly due to the influence of the Internet. Digital language includes a large number of discourse markers (DMs), many of which can be considered innovative. Acronymization, pragmatic specialisation, and compensatory lexical innovation are the most common lexical processes that can be witnessed in the DMs used in computer-mediated communication (CMC). The following novel DMs were identified in recent Twitter chats*: lol, tbh, omg, meh*, and *idk*. These DMs perform several functions, such as showing emotions, signaling uncertainty, hesitation, or mitigation. Interpreting these functions may not be an easy or obvious task for AI. The primary aim of the study is to evaluate the pragmatic competence of an LLM, Gemini 3 Pro, regarding the interpretation of these novel DMs. A mixed-method research process was employed: LLM-generated outputs were compared with the findings of the relevant literature, quantitative corpus analysis, and our qualitative human interpretation to assess the model's analytical usefulness. Gemini 3 Pro was found to show a high level of pragmatic competence in terms of interpreting the functions of DMs, but sometimes tended to overgeneralise, or failed to understand the tone of the text and the intention of the speaker to use a DM.

Keywords: discourse markers, LLMs, digital communication

1. Introduction, Research Objectives

Research on lexical innovation has become even more relevant in 21st-century society, particularly in the context of digital development, which is inevitably linked to multiple linguistic changes as well, including lexical innovations. Linguistic innovation is especially observable on social media platforms, whose discourse reaches broad audiences in a digital format that has become a norm in contemporary communication.

As Large Language Models (henceforth: LLMs) become more powerful and capable, it is now necessary to assess them beyond basic knowledge and data recall and focus on their ability to grasp nuance and context. In this pilot study, we focus on recent colloquial lexical innovations spreading on social media with discourse marking functions. In our pilot, we identified the following common discourse markers used in recent Twitter chats*: lol, tbh, omg, meh*, *idk*. These novel discourse markers (henceforth: DMs) are innovative tools that perform complex social tasks, such as showing emotions (such as *omg* showing surprise), signaling uncertainty and hesitation (by using *idk*), or ending a conversation in a text-based, digital environment (e.g. *so yeah*).

The primary aim of the study is to evaluate the pragmatic competence of an LLM (in our case study, Gemini 3 Pro) regarding novel discourse markers in computer-mediated communication (henceforth: CMC). (Gemini 3 Pro was released in November, 2025, described by Google as its most intelligent AI model yet, featuring enhanced reasoning and coding capabilities.) The secondary goal of the research is to compare and contrast the output of the LLM with human analysis, with the relevant findings of previous research on these items, as well as our own qualitative analysis of these items in our corpus.

2. Theoretical Background

2.1. Lexical Innovation

Social media remains the driving tool for versatile communication, reaching a large number of people and covering a wide variety of topics, including political, social, economic, and other information. It is also important to pay attention to the pragmatic side of such communication because of the way these topics are presented in the media, as the society's perception of social media discourse also depends on ideology, public acceptance, established stereotypes, accepted morality, gender perception, etc. (Pohorila, 2022). The dynamic phenomenon of modern media discourse also carries a powerful pragmatic and evaluative potential, influencing the views of society and forming subjective worldviews. In turn, research related to linguistic phenomena related to media discourse sheds light on linguistic form and corresponding genres guiding the pragmatic functions of social media discourses (Horbatko, 2021).

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 53–59

Lexical innovation has been studied from various perspectives by linguists, categorizing the processes of word formation (Miller 2014) and tracking the change of the meaning of words over time (Geeraerts 2010). In recent studies, researchers have focused on word lexicalization and their gradual acquisition of particular forms and meanings as well as the institutionalization of words as they enter into the standard vocabulary of a language (Brinton & Traugott, 2005). Lexicographical research focuses on identifying and defining neologisms by extensively using the corpus approach and internet search engine results (Kerremans et al. 2011). Traditional linguistic interest in the formation and development of new words is enhanced by corpus linguistics, which offers new solutions, as it allows for the open-ended analysis of language variation and change by searching large amounts of natural language data (Szmrecsanyi 2011; Grieve 2015). It should be admitted that small, regionalised corpora are not extensive enough to observe variations both in the use of common content words and rare new words, but the growth of social media has been changing the situation with the possibilities of computational linguists and LLMs to analyze incredibly large amounts of linguistic data harvested online, especially from Twitter, to understand certain patterns of lexical variation and change (Huang et al. 2016).

2.2. The Category of DMs

There is no widespread agreement in the literature about the name and the definition of this group of lexical items (including *tbh, omg, lol, idk, meh, so yeah*). Concerning the terminology of the present research, we refer to these items as discourse markers (DMs). DMs are traditionally defined as “sequentially dependent elements that bracket units of talk” (Schiffrin 1987: 31) or metalinguistic items that provide information about the segmentation and operation of a discourse (Fraser 1999). Schiffrin (1987) describes the role of DMs as “providing contextual coordinates for ongoing talk” that indicate for the hearer how an utterance is to be interpreted. This is the reason why they prove to be frequent and useful elements in CMC as well, since they help the readers disambiguate the intended meaning and tell us about the mental state/stance of the speaker or writer. Furkó (2014) overviewed earlier DM research and presented the criteria for DM identification by describing the key features of this morphologically, syntactically, and pragmatically heterogeneous group: high oral frequency, optionality (in a syntactic sense), low propositional contribution, procedural meaning, extreme multifunctionality (fulfilling pragmatic/interactional functions, such as stance, alignment, and mitigation), and context dependence. In summary, DMs are multifunctional pragmatic elements expressing various metacommunicative and cognitive functions. These functions of digital DMs will be explored in this research based on a small corpus and its mixed-method analysis, including analysis by an LLM.

2.3. Types of Lexical Innovation in DMs

We can classify the scrutinised DMs in terms of the type of lexical innovation present in them. The first feature is an abbreviation, as due to lexical economy, DMs shift from full phrases to initialisms. Such acronyms also turn into discourse markers with certain pragmatic functions. Another feature is pragmaticalisation, which means that acronyms undergo a lexicalisation process through which they become lexical units; for example, *lol* is used as a verb, e.g., in "*He literally lol'd*". Depending on the theoretical framework, scholars describe this process as pragmaticalisation (Ariel 1998), grammaticalisation (Traugott 1995), or constructionalisation (Traugott & Trousdale 2013).

2.4. Previous Research on the Selected Items in CMC

McCulloch (2019) describes language use on the Internet and argues that the English language is changing faster than before because of the influence of the Internet. Concerning the DMs under scrutiny, he finds that *lol* has become a softener. Using it at the end of a sentence (e.g., "*I'm so tired lol*") signals that the speaker is not actually complaining aggressively but rather seeking sympathetic feedback (e.g., a nod) from the reader.

Scott (2015) specifically analysed the one-to-many, asynchronous communication mode of Twitter and found that tweeters make their intended contextual assumptions accessible to a wide range of readers by using hashtags, which facilitate the use of an informal, casual style that fits the discourse context of Twitter. Scott suggests that expressions such as *tbh (to be honest)* and *ngl (not gonna lie)* serve as mitigators that soften the impact of a statement. These markers are typically used before giving potentially offensive or controversial opinions. Through the explicit reference to being honest, the writer signals a transition from polite conversation to a more authentic personal insight, which builds a sense of closeness with the listener (Scott 2015).

Tagliamonte & Derek (2008) analysed pragmatic particles in instant messaging among teens and found that *lol* and *omg* are used for discourse-pragmatic purposes. For instance, in their understanding, *lol* is used as a marker of empathy or a way to signal that the conversation is friendly. It serves a phatic function; it keeps the social connection open, rather than indicating actual humor. They claimed that teenagers use nonstandard language, but it should not be considered a degradation of language, but a new, innovative, and functional form of communication.

Vandekerckhove (2025) gave a detailed analysis of the use of *omg* in the digital language of Flemish adolescents. In general, he also shares the view that discourse markers function as

pragmatic signals that tell the reader exactly how to interpret the text, since digital writing lacks the nonverbal cues of face-to-face interpersonal communication. The paper highlights that, besides the primary function of *omg* to express shock, it fulfils discourse organizational functions as well. *omg* is often used to mark boundaries and bracket a message. It may signal a shift, e.g., from casual chat to a high-intensity narrative (as in *"omg you won't believe what happened..."*).

Softener markers like *lol* or *idk* are often used to mitigate face-threatening pragmatic acts. They are used when a speaker makes a request or expresses a slight criticism, and adding a marker at the end reduces the social risk of the interaction.

3. Methodology

We followed a mixed-method research process: a combination of quantitative corpus analysis (using a concordance), qualitative human interpretation, and AI-assisted analysis. LLM-generated outputs were compared with the findings of the relevant literature and our qualitative human interpretation to assess the model's analytical usefulness, pragmatic competence, and limitations, in line with recent computational studies on machine-learning methods for detecting hedges (Wise & El Barj 2023, p. 3). First, data was collected; a Twitter chat corpus was selected for this purpose, since it is a genre peculiar to recent digital communication. This text corpus was scraped from Twitter (242,170 words, 51 MB), where the odd lines are tweets and even lines are corresponding responded tweets. The corpus is formatted as a list of independent messages or short exchanges organised into one message per line. The text displays lexical variety, which is typical of social media, and includes a mix of standard English and slang (e.g., "*deadass*" and "*lowkey*"). The text is also highly informal, characteristic of non-standard capitalization, excessive punctuation (e.g., "*!!!!*"), and frequent use of emojis. In the subsequent stage of the experiment, AntConc 4.3.1, a freeware corpus analysis toolkit, was used for concordancing and quantitative text analysis (Anthony, 2024). We first carried out the quantitative and qualitative analysis of the pilot corpus and then prompted Gemini to carry out an LLM-based analysis. Finally, the different results were contrasted with one another.

First of all, we used AntConc 4.3.1, a freeware corpus analysis toolkit for concordancing and quantitative text analysis (Anthony, 2024). Most entries are single tweets or short chat messages, typically ranging from 10 to 25 words per line. Given the informal nature of the text, the TTR is relatively low for the entire corpus due to the repetition of common conversational phrases, though it contains a high number of unique informal variations and misspellings. The case-insensitive concordance searches of the pilot corpus gave the following counts for the target expressions: there are 201 occurrences of *lol* (including variations such as "*LOL", "lolol", "lolz"*), 35 occurrences of *tbh*, 21 *idk*, 15 *omg* and 4 *meh* items in our research corpus.

In the next stage, we discussed and agreed on the qualitative interpretation of the relevant lines (the left and right contexts of the word searches described above) as well as the functions of the digital DMs, driven by earlier works and the actual examples in the Twitter corpus. These findings were manually saved in a shared spreadsheet file for subsequent comparison.

As a next step, Gemini 3 Pro, an LLM-based generative AI chatbot, was used. Gemini 3 Pro was released in November 2025, described by Google as its most intelligent AI model yet, featuring enhanced reasoning and coding capabilities. In our experiment, Gemini was fed the full corpus file and was given the following prompt in thinking mode: "Collect and analyse the expressions *meh*, *lol, tbh, omg, idk* in the text file (Twitter chat corpus), analyse their uses, and classify the pragmatic and discourse functions of these expressions in digital communication." The prompt did not contain examples of classification or explicit category definitions. Subsequently, Gemini provided its answer about the common discourse-pragmatic functions and uses of the selected DMs (*mitigation, intensifying, hedging, expressing emotions, marking stance,* and *replacing facial expressions or gestures*), and it also reflected on usage in terms of the typical positions of the DMs, although it was not explicitly asked to do so. In the end of its reply, it also suggested giving an example for each function, if we need it. Therefore, we prompted in reply to: "give an example sentence for each function from the same text corpus attached." As a result, it gave a list of functions and an example for each, supposedly the most common usages (in the most frequent positions).

Two other LLMs were consulted and were given the same corpus file and prompt, but GPT-5.2 Auto by OpenAI and Claude 4.5 Sonnet did not upload the corpus file and did not perform the task, as the size of the attached file must have been too large. Claude 4.5 specifically highlighted that files larger than 31 mb cannot be uploaded, so these LLMs were not involved in the analysis, which has left the experiment as a single-model pilot study rather than a comprehensive comparative research study.

4. Research Findings

4.1 Human Interpretation: The Discourse-Pragmatic Functions of the Novel DMs in CMC

The utterances (in Table 1) demonstrate the typical functions and different positions of the scrutinised novel DMs: *lol, tbh, omg*, *meh* (expressing various emotions), and *idk* (expressing a hedge and functioning as a mitigator). These examples have been selected

by us from our corpus to illustrate the most common uses of the DMs in digital communication.

Marker	Example 1	Example 2
meh	*"A: Are you excited? B:* ***Meh****, not really."* (indifference)	*"The food was okay, but the service was just...* ***meh****."* (dissatisfaction)
lol	*"I can't believe I just sent that to the wrong person* ***lol****."* (mitigation)	*"****Lol****, that is literally the funniest thing I've seen today."* (irony)
tbh	*"I think the first season was better* ***tbh****."* (stance marking)	*"****Tbh****, I never really understood why that show was popular."* (hedging)
idk	*"****Idk*** *how females fuck with this."* (preface)	*"****Idk,*** *I'm just trying to help here"* (mitigation).
omg	*"****omg*** *same, I'm dying, it's all I've been thinking about"* (marks emotional state)	*"****OMG****, my phone has been jumping from like 39 to 0 if I open a new app; I'm fed up."* (intensifier)

Table 1. Examples from our corpus, complemented with human coding of the functions in brackets

4.2 Contrasting Human Interpretation and LLM Analysis

Analysis of "*meh*"

According to Urban Dictionary, *meh* represents doubt and functions as a shoulder shrug, with its meaning described in the entry:https://www.urbandictionary.com/define.php ?term=meh. In the Urban Dictionary it is not explicitly categorised as a DM, but in many contexts, we consider it a DM due to its low propositional contribution, procedural meaning and multifunctionality. In the tweets, it often expresses indifference and an evaluative stance as well, carrying an expression of dissatisfaction (in contrast with previous higher expectations), as in *"A: Are you excited? B: Meh, not really" in our chat corpus.* This signals that the writer or speaker finds the topic uninteresting and shows discouragement of further deep engagement on that specific point.

The LLM analysis does not always comply with reality, as *meh* is not always used initially, as described by the LLM. In fact, *meh* is common in mid-position as well in the research dataset. Moreover, it does not always serve as a DM (as was suggested by Gemini); sometimes it serves as a predicative adjective, as in "The *new update is a bit meh, I expected more features,"* where *meh* is remodified by *"a bit,"* and it means something like disappointing or unimpressive. The online version of the Cambridge Dictionary also mentions its use as an adjective.

Analysis of "*lol*" (laughing out loud)

Theoretically, this DM most commonly expresses the pragmatic function of mitigation. In our corpus, *lol* rarely indicates its literal meaning, laughter; instead, it more frequently functions as a mitigation in order to soften the blow of a critique. It is often used to signal irony or to suggest a friendly or non-confrontational tone.

Gemini analysis emphasizes the final position of this DM; however, we found several examples in our corpus where *lol* is used initially to indicate laughter, such as in "*lol that is literally the funniest thing I've seen today.*" In fact, it can be placed both at the beginning or at the end of a sentence (e.g., *"I'm dead, not looking forward to this lol"*).

Analysis of "*tbh*" (to be honest)

The main functions of this DM include stance marking and hedging. Concerning the hedging function, it is used for signaling an opinion that might be unpopular or controversial. It helps manage the user's face by showing that the statement is a subjective observation rather than an objective fact. It often appears at the beginning of a turn to frame the entire message as a moment of sincerity.

Gemini highlighted the sentence-initiality of *tbh*, but in fact, in our corpus of tweets, *tbh* was often placed on the left periphery (at the end) of sentences, as in the following examples: "*I'm just really tired of the constant drama, tbh*." "*I think the first season was better tbh*."

Analysis of "*idk*" (I don't know)

It usually carries a function of an epistemic hedge or mitigator used to signal uncertainty, a lack of commitment to a statement, or to soften the impact of a potentially controversial opinion. In this specific corpus, *idk* is pragmatically used to soften a potentially controversial opinion. It allows the speaker to simultaneously distance themselves from being expressively certain. For example**:** "idk how females fuck with this ". In this sentence, it functions as a preface to an observation, showing a personal confusion rather than an attack.

Concerning the functions of this DM, Gemini provided an analysis similar to human interpretation.

Analysis of "*omg*" (oh my god)

As indicated in the scientific literature, in our corpus *omg* signals high emotional arousal, expressing various emotions, such as surprise, shock, frustration, or excitement. It often serves to invite the interlocutor to share in their emotional state, as in *"omg, same. I'm dying, it's all I've been thinking about."* In this context, it expresses enthusiasm and acts as an intensifier for agreement.

Overall, the LLM, Gemini, provided quite an extensive pragmatic analysis in the context of digital communication (specifically Twitter), and its findings were most of the time in line with our own analysis. It managed to illustrate that the analysed expressions function less and less as literal semantic units and acquire the status of discourse markers indicating stance, which is a sign of the lexical innovation in the items. However, it included a few generalisations about the initial position of the DMs and the tone they express (e.g., *lol* expressing laughter, which was not always the case).

4.3 The Uses and Functions of Lexical Innovation in Digital Communication

Initialism and acronymization are the most visible forms of innovation representing the shift from full phrases to initialisms (*LOL, TBH, IDK, OMG*). The benefits and functions of the discussed innovative markers in digital media include speed, linguistic economy, and subcultural signalling. The innovation of lexical economy is driven by the principle of least effort. For example: *"Wait, you actually did that? lol stop."* Lexicalization demonstrates that the discussed acronymic DMs do not remain just short forms; they have become lexical units in their own right, as discussed above, *lol* being used as a verb (in *"He literally lol'd"*).

Obviously, digital communication lacks paralinguistic cues, such as tone of voice and facial expressions. Compensatory lexical innovation fills this gap with items such as *meh* or *omg* as innovative ways to represent a facial expression or a tone of voice. For instance, the DM *omg* represents innovative ways to display emotional intensity without audio or visual signals, and the DM *meh* functions as a translation of physical sounds into lexical entries. The transliteration of nonverbal cues clearly adds to lexical innovation, as *meh* represents an innovation where a nonlexical sound of a grunt is turned into a written word.

Lexical innovation may also involve pragmatic specialisation, which often involves a word becoming specialised for a specific social function. The best example is *tbh* used as a stance marker, which is almost always placed at the start of a sentence to affect the entire following message, to manage face, and to keep social politeness in a public or semi-public forum, such as Twitter.

5. Discussion about the Usability of LLMs in Discourse Analysis

The current experiment aligns with the study by Furkó (2025), stating that LLMs show the ability to identify common discourse markers and their functions. Our pilot experiment with Gemini was restricted, but it showed that the model proved surprisingly capable of explaining how new DMs function in digital conversations, often aligning well with established pragmatic theories. This suggests that AI could be an excellent help for researchers, sifting through large amounts of data to identify DMs and their environment for deeper human analysis. However, the study by Furkó (2025) shows that the AI analysis still has its limits. It tends to struggle with the subtler side of language, particularly when the meaning depends heavily on context. We saw a similar pattern in our own trial, as Gemini frequently missed the tone in our corpus and interpreted sarcastic comments as completely literal. This underscores a deficiency in pragmatic competence, because LLMs lack the situational awareness needed to safely make decisions about language appropriateness, e.g., if someone is being sincere, joking, or being sarcastic. Generally, AI sometimes still fails to grasp tone or irony because of the lack of real-world experience.

Additionally, in the current experiment, Gemini occasionally made broad assumptions about where DMs usually appear and let its own biases colour the findings in order to illustrate the initial usage of DMs.

6. Conclusion

Based on the analysis of our chat corpus, it was found that digital language includes a large number of DMs, many of which can be considered innovative. Acronymization, pragmatic specialisation, and compensatory lexical innovation are the most common lexical processes in CMC. The scrutinised digital DMs represent a class of true lexical neologisms, as they are built through compensatory innovation, which means that while tweets lack nonverbal cues such as tone of voice or facial expressions, a lexical innovation fills this gap. Words such as *meh* or the repetitive use of *omg* serve as innovative digital body language to replace a facial expression or a tone of voice that would otherwise be lost in text.

Gemini 3 shows a high level of pragmatic competence in terms of interpreting the functions of DMs but sometimes tends to overgeneralise (e.g. about the sentence-initial position of certain DMs). Gemini provides general linguistic rules about the position of DMs and simply does not mention the cases where DMs are used in uncommon positions, but we do not consider this generalisation a hallucination since the initial position is typical of DMs.

In brief, our pilot study indicates that an effective analytic approach may involve using AI tools to process large amounts of data to establish patterns, but it must be followed by critical human interpretation, especially when speakers/writers express sarcasm or irony or when semantic ambiguity is present in the text. Text generation by LLMs is not really genuine reasoning, though, and what is even more challenging for AI is to replicate pragmatic performance. While the scalability of AI is incredibly promising for linguistics, it is clear that a human analyst is still required to verify if the AI's interpretations are valid in a specific context.

7. Limitations

The limitation of the present pilot study is due to the lack of model diversity and the limited range of the analyzed data. Although we intended to include multiple LLMs, technical constraints regarding file size prevented GPT-5.2 Auto and Claude 4.5 Sonnet from performing the task, leaving the experiment as a single-model case study rather than a comprehensive comparative evaluation. The findings about the functions of these DMs in Twitter chats are in line with the functional categories of the related literature pertaining to computer-mediated communication. However, because the research is conducted on a relatively small dataset, consisting only of Twitter chats, the findings may not be fully representative of broader digital communication patterns. Besides, a monolingual approach was used. Our future research plans include testing further novel discourse markers (including multiword DM patterns as well) in multilingual and more extensive datasets.

8. Acknowledgments

This publication is based upon work from COST Action CA23147 GOBLIN—Global Network on Large-Scale, Cross-domain, and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

9. Bibliographical References

Ariel, M. 1998. Discourse Markers and Form-Function Correlations. In: Jucker, A. H. & Ziv, Y.. (Eds.) *Discourse markers: descriptions and theory*. Pragmatics and Beyond Series, 57. Amsterdam and Philadelphia: John Benjamins.

Brinton, L. and Traugott, E. 2005. *Lexicalization and language change*. Cambridge, UK: Cambridge University Press.

Fraser, B. 1999. What are discourse markers? *Journal of Pragmatics* 31, 931–952.

Furkó, P. 2014. Cooptation over grammaticalization. *Argumentum* 10, 289-300.

Furkó, P. 2025. Pragmatic markers and ideological positioning in EUROPARL: A corpus-based study. *Russian Journal of Linguistics* 29 (4). 795–816.

Geeraerts, D. 2010. *Theories of lexical semantics*. Oxford, UK: Oxford University Press.

Grieve, J., Nini, A. and Guo, D. 2017. Analyzing lexical emergence in American English online. *English Language and Linguistics* 21(1). 99-127.

Horbatko, A. O. 2021. Approaches to the definition of media discourse in modern English-language mass media. *Current issues of philology and methodology* (36-42). Sumy: Publishing and Printing Enterprise Printing Factory LLC.

Huang, Y., Guo, D., Kasakoff, A. & Grieve, J. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59. 244-255.

Kerremans, D., Stegmayr, S. and Schmid, H. 2011. The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In Kathryn Allan & Justyna Robinson (eds.), *Current methods in historical semantics*, 59-96. Berlin: Mouton de Gruyter.

McCulloch, G. 2019. *Because Internet: Understanding the New Rules of Language.* Riverhead Books.

Miller, G. 2014. *Lexicogenesis*. Oxford, UK: Oxford University Press.

Pohorila, A. I. 2022. The functioning of euphemisms in the English media discourse. *Transcarpathian Philological Studies*, 21(2), 100-103.

Schiffrin, D. 1987. *Discourse Markers.* Cambridge: Cambridge University Press.

Scott, K. 2015. The pragmatics of hashtags: Inference and conversational style on Twitter. *Journal of Pragmatics.* 81. 10.1016/j.pragma.2015.03.015.

Szmrecsanyi, B. 2011. Corpus-based dialectometry: A methodological sketch. *Corpora* 6(1). 45-76.

Tagliamonte, S. and Derek, D. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* - AMER SPEECH. 83. 3-34.

Traugott, E. G. 1995. The Role of the Development of Discourse Markers in a Theory of Grammaticalization. *Paper given at the 12th International Conference on Historical Linguistics.* Manchester; 13–18, August, 1995.

Traugott, E. and Trousdale, G. 2013. *Constructionalization and constructional changes*. Oxford: Oxford University Press. pp. 304.

Vandekerckhove, R. 2025. "OMG! Why discourse markers thrive in interactive social media writing" In: Fábián, A. & Trost, I. (eds.) *Impulses and Approaches to Computer-Mediated Communication: Proceedings of the 12th International Conference on Computer-Mediated Communication and Social Media Corpora.* University of Bayreuth.

Wise, M. & El Barj, H. N. 2023. PragMaBERT: Analyzing pragmatic markers in political speech. *CS224N Project Report.* Stanford University.

10. Language Resource References

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc (accessed on 6 October, 2025)

Anthropic. (2025). *Claude 4.5 Sonnet* [Large language model]. https://claude.ai/ (accessed on 27 December, 2025)

Google. (2025). *Gemini 3 Pro* [Large language model]. https://gemini.google.com/ and https://aistudio.google.com/ (accessed on 28 December, 2025)

Marsan's Twitter chat corpus repository: https://github.com/marsan-ma/chat_corpus/blob/master/twitter_en.txt.gz, subpart of this dataset: https://github.com/marsan-ma/chat_corpus (accessed on 5 November, 2025)

"Meh" in *Cambridge Dictionary*: https://dictionary.cambridge.org/dictionary/english/meh (accessed on 15 November, 2025)

"Meh" in *Urban Dictionary*: https://www.urbandictionary.com/define.php?term=meh (accessed on 15 November, 2025)

OpenAI, Inc. (n.d.). Models. *GPT-5.2.* [Large language model]. Platform.openai.com; OpenAI, Inc. https://platform.openai.com/docs/models/ (accessed on 28 December, 2025)

A Comparative Evaluation of Semantic Ambiguity Detection in Two LLMs

Lili Tamás

Károli Gáspár University of the Reformed Church in Hungary
Reviczky u. 4. 1088 Budapest, Hungary
lili.tamas.contact@gmail.com

Abstract

The growing popularity and misconceptions about conversational AI systems are driving efforts to establish a universally accepted framework for evaluating large language models. Testing large language models on tasks designed to assess human cognitive skills has become widespread. This paper presents the results of a pilot experiment and a comparative evaluation of the ability of OpenAI's GPT-4.1 and GPT-4.1 mini to detect semantic ambiguity based on the works of Shultz and Pilon (1973) and Zipke et al. (2009). The experiment used a task sheet of 116 items utilising riddles, single sentences, and sentence pairs. It included systematically varied instructions on a four-level scale ranging from no mention of ambiguity to direct mention. Lexical and structural ambiguity were both employed, including surface-structure and deep-structure ambiguity. The results suggest that even advanced models, such as GPT-4.1 and GPT-4.1 mini, tend to consider only one possible meaning of ambiguous sentences. However, the recognition of ambiguity improved quickly when the possibility of ambiguity was explicitly referenced in the instruction. Additionally, the results imply that model size is not directly connected to performance, as GPT-4.1 scored better on lexical ambiguity detection tasks, while GPT-4.1 mini surpassed the larger model in structural ambiguity detection.

Keywords: LLM evaluation, semantic ambiguity, contextual limitations, language modelling

1. Introduction

The rapid advancement of large language models imposes new challenges for performance evaluation. Compared to benchmarking in evaluating discriminative language models, assessing the performance of generative language models poses new challenges. As of now, there is no universally accepted framework for evaluating such models (Wolters et al., 2024; Tam et al., 2024; Seo et al., 2024; Miaschi et al., 2024). Methods that simulate tests that were initially intended to assess human cognitive skills have become widespread. One topic of interest is LLMs' ability to detect ambiguity and whether these systems demonstrate metalinguistic awareness. This report presents the results of a small-scale experiment and the comparative evaluation of the ability of OpenAI's GPT-4.1 and GPT-4.1 mini to detect semantic ambiguity.

Assessing whether large language models exhibit metalinguistic awareness is of utmost importance, as the general belief is that such models have a deep linguistic understanding of languages, beyond that of native speakers. As Rohr-Brackin (2025) explains, metalinguistic awareness is "a part of general cognition" (2025, p. 28). Metalinguistic awareness is the active attention to the knowledge domain "that describes the explicit properties of language" (Bialystok, 2021, qtd. in Roehr-Brackin, 2025, p. 28). Illiteracy and metalinguistic awareness are connected, as "illiterate adults' metalinguistic awareness remains at low levels, such as the ability to identify rhymes … despite cognitive maturity" (Roehr-Brackin, 2025, p. 29). Recognition of ambiguity requires metalinguistic awareness, and as such, it was chosen as the domain of the present research.

While the challenges of semantic ambiguity detection in NLP have been the focus of research for decades, Jayaweera and Dorr (2025) state that annotator-disagreement stemming from linguistic ambiguity is still often considered noise rather than a reflection of "meaningful, coexisting interpretations" (p. 37). Jayaweera and Dorr (2025) highlight that the "absence of gold-standard annotations for different ambiguity types hinders progress in training and evaluating models that aim to align more closely with human interpretive processes" (p. 44). The authors emphasise the "need for the creation of new datasets specifically annotated for ambiguity presence and type" and see "exploring unsupervised or weakly supervised methods" as promising (Jayaweera & Dorr, 2025, p. 44). The experiment presented in this current paper is to serve as a pilot project for a larger scale experiment as a step towards achieving the goals defined by Jayaweera and Dorr (2025).

2. Methods

The experiment tested the abilities of GPT-4.1 and GPT4.1 mini. OpenAI defines GPT-4.1 as their "smartest non-reasoning model", while GPT-4.1 mini is the "smaller, faster version of GPT-4.1" (OpenAI, Inc., n.d.). OpenAI does not share the detailed technical specifications of its models, but both models have a 1,047,576-token context window (OpenAI, Inc., n.d.). The tasks

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 60–65
16 MAY 2026.

used in the experiment were based on Shultz and Pilon's (1973) and Zipke et al.'s (2009) work on testing children's ability to detect and comprehend linguistic ambiguity. Each task consisted of a context and an instruction. As shown in Table 1, three types of contexts were used.

Context ID	Original study	Task type	Ambiguity type	Context
S5	Zipke et al. (2009)	Single sentence	Deep-structure	Flying kites can be exciting.
SP9	Shultz and Pilon (1976)	Sentence pair	Surface-structure	She helped the boy with the hat. She helped the boy put on his hat.
R8	Zipke et al. (2009)	Riddle	Lexical	Why is a school yard larger at recess than at any other time? a. At recess there are more feet in it. b. It isn't.

Table 1: Context type examples

The first type was riddles in which the humour comes from lexical ambiguity. The second type was single, either lexically or structurally ambiguous sentences, including surface-structure and deep-structure ambiguity. The third type of context was sentence pairs in which the first sentence was ambiguous, and the second sentence unambiguously conveyed one of the possible meanings of the first sentence. These sentence pairs also included lexically and structurally ambiguous sentences, incorporating both surface-structure and deep-structure ambiguity.

Kess and Hoppe (1981) define lexical ambiguity as the result "of a word or word sequence having more than one distinct meaning" (p. 30). Surface structure ambiguity "reflects two distinct syntactic groupings of adjacent words in the string … Deep structure ambiguity, on the other hand, reflects different logical relational sets between words or phrases in the sentence." (Kess & Hoppe, 1981, p. 31). The authors point out that in the experiment of Mackay and Bever (1967), the participants "the median perception time for the detection of ambiguities went from lexical to surface structure to deep underlying structure ambiguities", suggesting differences in the difficulty of recognising these three types of ambiguities (Kess & Hoppe, 1981, p. 31).

In this current experiment, there were eight riddles and seven single sentences by Zipke et al. (2009) and 14 sentence pairs by Shultz and Pilon (1973). This resulted in 29 context texts overall. These 29 context texts were used across four task sheets, for a total of 116 tasks per model. The four task sheets (Level 1-Level 4) differed in the style of instructions. Table 2 below contains all instructions based on context type over the four levels.

	Riddles	Single sentences	Sentence pairs
Level 1	Please choose the correct answer to the question and explain your decision.	Please explain the meaning of the sentence.	Please explain the meaning of both sentences.
Level 2	Please choose the correct answer to the riddle and explain your decision. Please also explain why the riddle is funny.	Please explain the possible meanings of the sentence.	Please explain the meaning of both sentences. Please also compare the two sentences.
Level 3	Please explain what makes this a riddle and why it is funny.	Please explain all possible meanings for the sentence.	Please explain all possible meanings for both sentences.
Level 4	Ambiguity makes this riddle funny. Both answers can be considered true. Please explain why. Please also choose which answer is actually true.	The sentence is ambiguous. Please explain all possible meanings for the sentence.	One of the sentences is ambiguous. Which one? Please also explain all possible meanings for both sentences.

Table 2: Instructions for the different levels

As shown, in the case of the riddles, the instructions for Level 2 and Level 3 experimented with ways to encourage the models to analyse the riddles, then contained a direct mention of ambiguity on Level 4. In the case of the single sentences, the instructions contained no reference to ambiguity on Level 1, a hint at ambiguity on Level 2, a more explicit hint on Level 3, and directly referenced ambiguity in Level 4. For the sentence pairs, a combination of encouraging analysis and a direct mention of ambiguity on Level 4 was used. Model responses were elicited via API calls with default parameter settings, which OpenAI do not specify in the available documentation.

3. Results and discussion

The answers of the models were assessed on a binary scale. On Level 1 to Level 3, the assessment was either “Noticed ambiguity” or “Didn’t notice ambiguity”. As on Level 4, ambiguity is addressed in the instruction, and answers were assessed as either “Correct interpretation” or “Incorrect interpretation”. Figure 1 below shows the percentage of correct answers by models and levels:

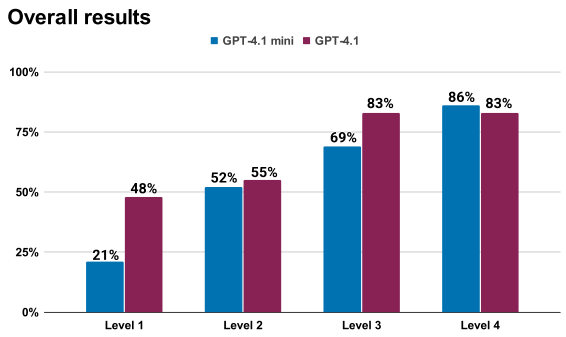

Figure 1: Percentage of correctly solved tasks

As the results show, without at least a hint at ambiguity or encouragement to analyse, the models tended to consider only one possible meaning. With the introduction of referencing the possibility of multiple meanings, both models’ results improved. Interestingly, the results of GPT-4.1 did not increase when the ambiguity was explicitly mentioned at Level 4 compared to Level 3, whereas the results of GPT-4.1 mini improved with every level. While GPT4.1 mini achieved a low score on Level 1, it surpassed GPT-4.1 when ambiguity was explicitly mentioned. Nonetheless, both models showcased significant improvement through more direct prompting.

The two models’ performance varied significantly depending on the type of ambiguity. Figure 2 shows the results for tasks utilising lexical ambiguity:

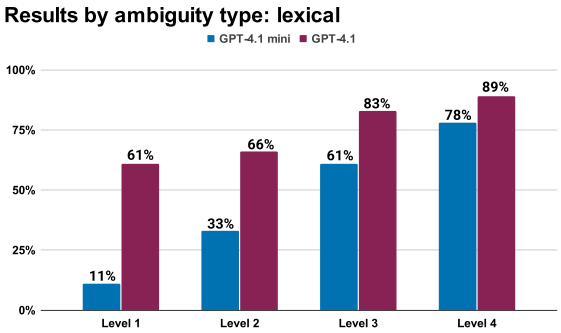

Figure 2: Results of lexical ambiguity tasks

GPT-4.1 consistently outdid GPT-4.1 mini when the task was lexical ambiguity detection. On Level 1, when the instruction does not reference possible double meanings in any way, GPT-4.1 correctly solved 61% of tasks, while GPT-4.1 mini scored only 11%. GPT-4.1 mini showed significant improvement throughout the levels, and GPT-4.1 also continued to steadily improve after the strong start. Nonetheless, as Figure 2 shows, the direct mention of ambiguity did not result in a perfect score for neither of the models.

While the tasks utilising lexical ambiguity seemingly show that the larger model, GPT-4.1, is superior, the results related to structural ambiguity provides a new perspective.

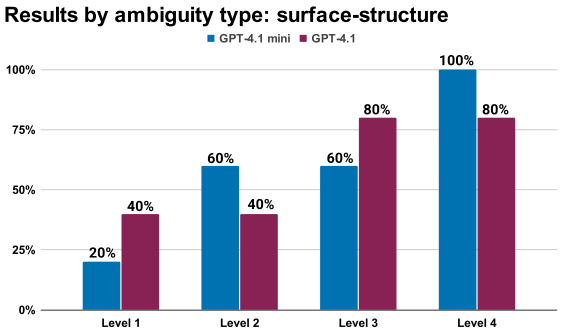

Figure 3: Results of surface-structure ambiguity tasks

As Figure 3 displays, GPT-4.1 mini surpassed GPT-4.1 on Levels 2 and 4. When there was no indication of possible ambiguity, GPT-4.1 outscored GPT-4.1 mini, but on Level 4, GPT-4.1 achieved a perfect score, while GPT-4.1 solved 80% of the tasks correctly. Levels 2 and 3 also highlight an interesting difference between the way instructions affect the models’ performance. GPT-4.1 improved. While GPT-4.1 mini scored better when the possibility of double meanings was offered, the difference between “the possible meanings” and “all possible meanings” did not result in a better performance. In contrast, for GPT-4.1, the difference in instructions between Level 2 and 3 mattered, but the direct mention of ambiguity did not between Levels 3 and 4.

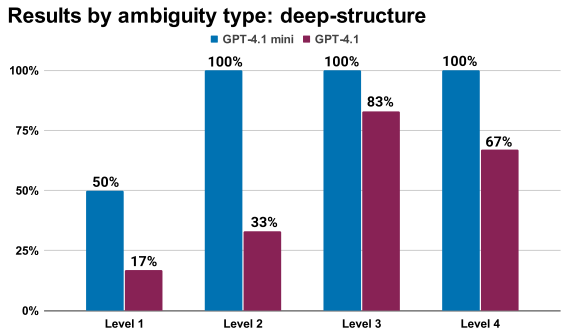

Figure 4: Results of deep-structure ambiguity tasks

Surprisingly, GPT-4.1 mini significantly outperformed GPT-4.1 in deep-structure ambiguity detection. Figure 4 shows that GPT-4.1 mini scored 50% percent on Level 1, and solved all tasks correctly on Levels 2-4, whereas GPT-4.1 scored only 17% percent on Level 1, and its performance peaked on Level 3, with a drop on Level 4.

These results indicate that while GPT-4.1 is better in detecting lexical ambiguity, GPT-4.1 mini surpasses the other model in structural ambiguity detection. This strongly suggest that a larger model size does not necessarily lead to better performance.

As expected, whenever the possibility of ambiguity was implied in the instructions, both models offered various options as facts in all cases, even when they failed to identify all meanings. One such example is related to task SP4 shown in Table 3 below. Neither of the two models considered "tank" as a vehicle but suggested that the gas was put into a gas cannister on various levels. On Level 4, GPT-4.1 mini even identified the second sentence, "He put some gas in the car" as the ambiguous one, and stated that the first sentence, "He put some gas in the tank" is not ambiguous.

Context ID	Original study	Task type	Ambiguity type	Context	Note
SP4	Shultz and Pilon (1976)	Sentence pair	Lexical	He put some gas in the tank. He put some gas in the car.	Both models failed all tasks using this context text. → Neither model considered "tank" as a vehicle, only as "gas tank."
SP11	Shultz and Pilon (1976)	Sentence pair	Deep-structure	It is really quite wonderful to see. It is really a wonderful sight.	Both models failed all tasks using this context text. → Neither model considered the meaning connected to a (lack of) visual impairment.

Table 3: Interesting examples from the results

The examined models' apparent difficulty to detect semantic ambiguity unprompted carries implications relevant to the study of neology. Neology often begins in an ambiguous zone, where a word or phrase is reused in a novel context, and this new use is initially ambiguous or inferable. Then, repeated contextual anchoring stabilises a new meaning, and ambiguity either persists (resulting in polysemy) or resolves (resulting in specialisation). Therefore, ambiguity could be seen as the transitional state of neology before the new meaning reaches high-enough frequency of use (Bybee, 2006).

Metaphorical use of a word or phrase can also result in neologisms (Bowdle & Gentner, 2005). Such examples are "virus" or "cloud" that acquired new, abstract meanings in computing, referencing the base concepts through similarity, resulting in lexically ambiguous words. Early uses were ambiguous and listeners relied on pragmatic inference in uses such as "store your files in the cloud" The sentence "The virus is spreading." remains ambiguous without additional context despite the new meaning having been lexicalised.

Future experiments testing LLMs' semantic ambiguity detection abilities could utilise neologisms, and the results could reveal frequency distribution between base concepts and target concepts. Additionally, LLMs could be tested on their ability to comprehend neologisms

in semantically ambiguous contexts. A possible prompt for an experiment/pilot study could be the following: “In the sentence ‘They left their data in the cloud,’ list all plausible interpretations and rank them by likelihood for usage in 2005, 2010, and 2020.” The answers could be compared to data from tools such as Google’s Books Ngram Viewer.

4. Conclusion

The quantitative evaluation of LLMs presents an ongoing challenge. To contribute to these efforts, this study examined OpenAI’s GPT-4.1 and GPT-4.1 mini models in terms of semantic ambiguity detection and comprehension. GPT-4.1 mini linearly improved with more direct instructions, while GPT-4.1 reached the same percentage of correct solutions on Level 3 and Level 4. The results show that without direct instructions even advanced models, such as GPT-4.1 and GPT-4.1 mini, tend to consider only one possible meaning, most likely based on word or phrase frequency. Additionally, this experiment revealed that GPT-4.1 scores better on tasks utilising lexical ambiguity, while GPT-4.1 mini outperformed the larger model in detecting structural ambiguity, implying that model size is not directly linked to performance. The findings of this experiment could be utilised in neology research, as neologies are oftentimes ambiguous, especially after emergence.

5. Limitations

The results of this study must be considered in the light of its many limitations. Nevertheless, while the experiment was small-scale, the results show that a larger-scale experiment would be beneficial.

Future research plans include building a larger task set with a more balanced inclusion of ambiguity types, as in the current task sheet lexical ambiguity is over-represented. A similar pilot study could be carried out on sentences including recently coined ambiguous neologisms. Furthermore, incorporating a human baseline (native and non-native speakers with varying levels of verified language proficiency) would allow direct comparison between human and LLM abilities in the semantic ambiguity detection domain. To create more reliable assessments, incorporating multiple runs and averaging the results as well as a more nuanced assessment scale are necessary for future experiments. After refining the experimental design, the goal is to test non-commercial as well as other commercially available models.

6. Bibliographical References

Bowdle, B. F., & Gentner, D. (2005). The Career of Metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295x.112.1.193

Bybee, J. L. (2006). From Usage to Grammar: The Mind’s Response to Repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Jayaweera, C., & Dorr, B. J. (2025). From Disagreement to Understanding: The Case for Ambiguity Detection in NLI. *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, 37–46. https://doi.org/10.18653/v1/2025.nlperspectives-1.4

Kess, J. F., & Hoppe, R. A. (1981). *Ambiguity in Psycholinguistics* (H. Parret & J. Verschueren, Eds.; pp. 1–123). John Benjamins Publishing.

Miaschi, A., Dell’Orletta, F., & Venturi, G. (2024). Evaluating Large Language Models via Linguistic Profiling. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2835–2848. https://doi.org/10.18653/v1/2024.emnlp-main.166

Roehr-Brackin, K. (2025). Measuring children’s metalinguistic awareness. *Language Teaching*, *58*(1), 27–43. https://doi.org/10.1017/s0261444824000016

Seo, J., Choi, D., Kim, T., Cha, W. C., Kim, M., Yoo, H., Oh, N., Yi, Y., Lee, K. H., & Choi, E. (2024). Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study. *Journal of Medical Internet Research*, *26*, e58329. https://doi.org/10.2196/58329

Shultz, T. R., & Pilon, R. (1973). Development of the Ability to Detect Linguistic Ambiguity. *Child Development*, *44*(4), 728. https://doi.org/10.2307/1127716

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01258-7

Wolters, A., Arz Von Straussenburg, A., & Riehle, D. (2024). *Evaluation Framework for Large Language Model-based Evaluation Framework for Large Language Model-based Conversational Agents Conversational Agents.* https://aisel.aisnet.org/pacis2024/track01_aibussoc/track01_aibussoc/14

Zipke, M., Ehri, L. C., & Cairns, H. S. (2009). Using Semantic Ambiguity Instruction to Improve Third Graders’ Metalinguistic Awareness and Reading Comprehension: An

Experimental Study. *Reading Research Quarterly*, *44*(3), 300–321. https://doi.org/10.1598/rrq.44.3.4

7. Language Resource References

OpenAI, Inc. (n.d.). Models. Platform.openai.com; OpenAI, Inc. Retrieved October 8, 2025, from https://platform.openai.com/docs/models/

Google. (n.d.). *Books Ngram Viewer*. Google.com. https://books.google.com/ngrams/

LLM-Based Frame and Stance Annotation for 19th-Century Rumour Discourse in US and UK Newspapers: A Digital Humanities Resource

Wanshu Zhang

University of Luxembourg
2, Av. de l'Universite
wanshu.zhang@uni.lu

Abstract

Digital humanities scholars increasingly consult digitized historical newspapers to study how rumours travel, how institutions respond, and how everyday publics negotiate credibility. Yet this interpretive work is slowed by two bottlenecks: noisy OCR text that obscures variant spellings and layouts, and the absence of scalable semantic annotations for what a rumour is "about" and how a text positions itself toward it (assertion, denial, attribution, correction). I present a resource-building pipeline that links a previously validated retrieval workflow for rumour discourse with a new evaluation setup for large language models as semantic annotators. From public-domain US and UK newspapers (1804–1896), I derive sentence-level rumour instances and construct an 800-instance balanced benchmark for (i) topical framing (7-way), (ii) evidential stance (4-way), and (iii) an optional audit flag for temporal anachronism in model rationales. I also report a preliminary pilot with Gemini 2.5 Flash-Lite on a 200-instance singly annotated subset, showing that structured JSON output is operationally stable and that evidential stance appears more tractable than topical framing under the current prompt design. The outcome is a transparent benchmark design and annotation protocol that can be extended to other periods and languages.

Keywords: digital humanities; historical newspapers; rumour discourse; semantic frames; evidentiality; large language models; evaluation

1. Motivation and digital humanities relevance

Rumours are not only false information; they are a historically situated genre through which communities register uncertainty, manage risk, and negotiate authority. In nineteenth-century newspapers, rumours cluster around wars and diplomacy, financial panics, epidemics, crime, and moral anxieties—topics that also structure the archival record that historians and literary scholars read today. Large-scale collections make it possible to trace these dynamics across decades, but the core DH challenge is interpretability: scholars need evidence for what is being talked about and how the text positions itself toward the circulating claim.

The project also speaks directly to lexical change and semantic evolution, since rumor-marking and evidential formulas such as it is said, we learn, and a correspondent writes shift in frequency and function across the nineteenth century.

My project contributes by treating LLMs as assistive annotators for rumour discourse: I aim to speed up the first pass of corpus exploration while preserving the ability to audit and contest model decisions. This paper describes (i) a benchmark derived from public-domain US and UK newspapers (1804–1896), (ii) an annotation framework for topical framing and evidential stance, and (iii) a pilot evaluation setup for LLM-assisted analysis that remains legible to humanists.

I align the annotation task with practical DH questions. For example: When do newspapers attribute rumours to named institutions versus anonymous correspondents? Do denials and corrections concentrate around specific topics? Are gossip-like rumours framed differently across regions and decades? Answering such questions requires structured semantics, but it also requires transparency about uncertainty and historical context.

2. Related work across DH and NLP

Historical newspaper analysis faces persistent obstacles: OCR errors, layout artefacts, non-standard orthography, and domain shift across time and publication venues. These issues motivate robust retrieval and cleaning methods, as well as evaluation practices that report uncertainty rather than hiding it. DH work has emphasized collaborative curation and interpretive workflows that combine computational signals with expert reading, rather than one-shot automation.

In rumour studies, historians have treated false or unverified news as a window into social psychology and information ecologies, from

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 66–70

wartime “false news” to moral panics. In NLP, stance detection and framing analysis are well studied for modern sources, but label sets and model behaviors often assume present-day discourse conventions and do not account for historically specific evidential formulas such as we learn, it is rumored, or a correspondent writes.

Modern stance-detection work typically targets contemporary discourse and assumes relatively stable present-day evidential conventions. By contrast, historical NLP must contend with OCR noise, orthographic variation, and temporal domain shift, while nineteenth-century newspapers also rely on period-specific formulas of attribution, hedging, and correction. My benchmark adapts stance-style annotation to this setting by combining a retrieval workflow validated in a DH venue with prompts that explicitly discourage presentist reasoning and require textual grounding.

3. Data and derived units

Source corpora include public-domain newspaper datasets for the US and the UK. The material spans 1804–1896 and covers diverse genres, including local news, international dispatches, editorials, and advertisements embedded in text streams. I treat the rumour instance as the core unit: a sentence, or short sentence group, that asserts, attributes, denies, or corrects a circulating claim.

Using an established two-phase workflow (structural OCR cleaning and orthography-robust retrieval), I identify candidate rumour sentences in the two corpora. Dependency patterns are used only at the retrieval stage to recover proposition-like instances despite historical variation and OCR noise. After light filtering, I remove severely corrupted OCR spans, exact or near-duplicates, and non-informative fragments lacking a recoverable rumour proposition. The benchmark is built through retrieval, sampling, and human annotation; LLMs do not create gold labels.

From this candidate pool, I construct a balanced benchmark sample for annotation and subsequent LLM evaluation. The current sample contains 800 instances, evenly divided by region (400 US, 400 UK) and distributed across mid-century time bins so as to capture major shifts in press infrastructures, including telegraphy and news agencies, without over-representing any single period. Because annotation is still ongoing, the present paper does not yet define a final train/dev/test partition; instead, it focuses on benchmark construction, annotation protocol, and an initial pilot subset of 200 annotated instances. The current pilot uses Gemini 2.5 Flash-Lite only.

4. Annotation targets and interpretive rationale

Each rumour instance is annotated along two axes: topical framing and evidential stance. Topical framing captures the primary social domain in which the rumour is presented, using seven labels: WAR_DIPLOMACY, MARKETS_COMMERCE, CRIME_JUSTICE, HEALTH_EPIDEMIC, POLITICS_PUBLIC_LIFE, SCIENCE_TECHNOLOGY, and SOCIAL_CULTURAL_LIFE. Evidential stance captures how the text positions the circulating claim, using four labels: ASSERTED, ATTRIBUTED, HEDGED, and DENIED_CORRECTED.

Each instance receives one label per axis. Annotators assign the topical-framing label that best matches the primary domain foregrounded in the passage rather than all potentially relevant themes. For evidential stance, ASSERTED is used when the claim is presented with minimal hedging, ATTRIBUTED when it is linked to a named or inferable external source, HEDGED when uncertainty or rumor-marking language is foregrounded without clear attribution, and DENIED_CORRECTED when a circulating claim is explicitly rejected, corrected, or countered.

To improve consistency, the guidelines prioritize local textual evidence over inferred background context. Formulaic expressions such as it is said, we hear, or we learn are treated as HEDGED unless the wording clearly attributes the claim to a specific source, in which case ATTRIBUTED is preferred. Likewise, passages that repeat a circulating claim only in order to reject it are labeled DENIED_CORRECTED rather than ASSERTED.

4.1 Mini-examples for label legibility

The protocol includes short representative examples to keep the labels intelligible for humanities readers and annotators. For evidential stance, ATTRIBUTED is illustrated by formulations such as “A dispatch from Vienna states that ...”; HEDGED by “It is rumored that cholera has appeared ...”; DENIED_CORRECTED by “The report of a bank failure is unfounded ...”; and ASSERTED by “The prisoner confessed ...”. These examples are not substitutes for historical reading, but scaffolding that helps annotators recognize period-appropriate evidential signals and avoid importing contemporary assumptions.

5. Human annotation protocol and adjudication

Because DH resources must be trustworthy and reusable, this paper specifies a lightweight but explicit human annotation protocol. Annotators receive (i) the cleaned rumour span, (ii) minimal metadata (year, region), and (iii) label definitions with decision rules. They do not receive full article context by default, in order to keep the unit comparable with model inputs; however, optional context lookup is permitted when OCR fragmentation or severe ambiguity makes the span difficult to interpret.

The benchmark is designed for full double annotation by two human annotators with backgrounds in digital history, including the author and a second annotator. Each instance receives one topical-framing label and one evidential-stance label. At the time of submission, 200 instances have been annotated by the author as an initial pilot subset, while full double annotation of the 800-instance benchmark is ongoing. After independent annotation, disagreements will be reviewed in adjudication sessions, with brief notes recorded for recurrent borderline cases. These notes are intended to form part of the released resource documentation and to make interpretive decisions transparent for future users.

Once the double-annotation pass is complete, I will compare the two annotators' decisions, revise guideline wording where necessary, and produce an adjudicated gold set for subsequent model evaluation. Because annotation is still in progress, I do not yet report inter-annotator agreement; instead, the present paper reports a preliminary model pilot on the singly annotated subset.

6. Pipeline overview and LLM prompting

I organize the project as a reproducible pipeline from corpus to benchmark construction, human annotation, and model evaluation. Phase A produces cleaned rumour candidates and structured metadata. Phase B evaluates ChatGPT, Claude, and Gemini as semantic annotators under historically informed prompting. My aim is not to construct an exhaustive leaderboard, but to compare how widely used general-purpose LLMs handle topical framing and evidential stance in nineteenth-century rumour discourse under controlled prompt conditions.

I test three prompt variants: P1 (minimal label definitions), P2 (definitions plus short historical examples), and P3 (definitions plus a caution against presentist reasoning and a requirement to cite local textual evidence). For the present pilot, I generate a single structured response per instance and retain full model outputs, including predicted labels and rationales, in order to support auditing and later re-analysis.

Each model takes as input a cleaned sentence span with minimal metadata (year, region), and produces as output a topical frame, a stance label, an optional audit flag, and a short justification grounded in quoted words or phrases. In the current submission, I use this setup for a preliminary pilot with Gemini 2.5 Flash-Lite on a singly annotated 200-instance subset. The pilot reported here uses the P3 prompt variant. For the current pilot, Gemini 2.5 Flash-Lite produces one structured response per instance, and pilot results are reported as diagnostic accuracy against the available single-annotator reference labels. Full comparative evaluation across models will follow once the adjudicated benchmark is complete.

7. Evaluation plan and DH-oriented reporting

I outline the full evaluation framework for ChatGPT, Claude, and Gemini once the adjudicated benchmark is complete. As an initial pilot, however, I run Gemini 2.5 Flash-Lite on a singly annotated subset of 200 instances in order to test prompt usability, inspect output rationales, and identify recurrent error types before full-scale comparative evaluation.

Standard metrics such as macro-F1 and confusion matrices remain necessary, but are insufficient for DH use on their own. I therefore also plan to report per-region and per-period breakdowns and audit-flag rates in order to capture temporal variation and unsupported or presentist rationalization.

Gemini returned valid structured JSON for all 200 cases, indicating that the prompt and output schema are operationally stable for batch annotation. On this preliminary subset, the model achieved 0.66 accuracy for topical framing and 0.745 accuracy for evidential stance when compared against the available single-annotator reference labels. The pilot used the P3 prompt variant and is intended as a diagnostic rather than benchmark-final evaluation, helping me test prompt stability, output structure, and recurrent confusion patterns before full adjudicated comparison. Because the current pilot subset is singly annotated and label distributions are uneven, I treat these scores as provisional evidence of task feasibility rather than as final performance claims.

In the pilot, topical framing is strongest for WAR_DIPLOMACY and MARKETS_COMMERCE, and weaker for broader categories such as SOCIAL_CULTURAL_LIFE. For stance, performance is strongest on the dominant

HEDGED class. Even so, the pilot is useful for validating the prompt design and identifying confusion patterns for later adjudicated evaluation.

I complement aggregate scores with qualitative error typologies that matter for interpretation. Recurrent categories include (i) collapsing WAR_DIPLOMACY into POLITICS_PUBLIC_LIFE when dispatches mention ministers or cabinet changes; (ii) confusing ATTRIBUTED with HEDGED in formulaic phrases such as it is said or we learn; and (iii) over-triggering HEALTH_EPIDEMIC for metaphorical uses of terms such as plague.

7.1 A DH use case: mapping credibility work

As an illustrative DH use case, I propose a credibility work map that combines topical framing with evidential stance to identify where newspapers perform verification, denial, or distancing. For example, a spike in DENIED_CORRECTED within MARKETS_COMMERCE during a financial panic can be read alongside editorials about speculation and information flows. Likewise, persistent ATTRIBUTED stance within WAR_DIPLOMACY may highlight reliance on telegraphed dispatches and named correspondents.

8. Resource plan (LRE Map)

I plan to release the benchmark and protocol as a DH-facing language resource with clear documentation. The release package will separate (a) benchmark text spans and metadata, (b) human annotations and adjudication notes, and (c) model outputs produced under documented prompt conditions. It will also include normalization notes, dependency-pattern templates used in retrieval, label definitions, and examples for ambiguous cases, so that the resource remains reusable beyond a single model or corpus.

9. Conclusion

I have described a DH-grounded resource-building pipeline that connects robust rumour retrieval in historical newspapers with an annotation framework and an evaluation design for LLM-based topical framing and evidential stance analysis. Rather than presenting a completed benchmark, the paper introduces a concrete protocol for constructing one in a transparent and historically sensitive way. The preliminary Gemini pilot shows that structured model output is operationally feasible, while also highlighting the need for fuller annotation and adjudication.

Next steps include completing double annotation and adjudication for the balanced 800-instance benchmark, releasing prompt templates and evaluation scripts, extending the pilot to additional models, and testing the resource with DH researchers.

10. Limitations and ethical considerations

Historical newspapers contain sensitive material, including racialized language, moral panics, and stigmatizing descriptions that may be reproduced or amplified by automated tools. I therefore recommend that downstream users treat the benchmark as an index for scholarly inquiry rather than as a stand-alone truth source, and that they document interpretive choices when using model annotations in publications. When releasing model outputs, I will include guidance on handling harmful content and on avoiding over-interpretation of noisy OCR spans.

Methodologically, the scheme is intentionally compact in order to suit a short paper and a first resource release; it does not yet capture richer discourse structure such as multi-step attribution chains, editorial genre differences, or networked rumour propagation across titles.

11. Reproducibility and planned release

To make the resource reusable for DH audiences, I will publish a compact “how to cite and reuse” guide alongside the benchmark, including provenance fields, a recommended citation, and a changelog. I will also release prompt templates and evaluation code so that other researchers can rerun the annotation experiment with different models or local systems. Because reproducibility in historical corpora is complicated by OCR post-processing and collection updates, I store both the cleaned span used for annotation and a pointer back to the raw source text whenever stable identifiers are available; otherwise, I provide hashed text fingerprints for alignment across releases.

12. Illustrative micro-reading: how the labels support interpretation

To demonstrate how the scheme supports interpretation rather than purely technical classification, I outline a simple micro-reading workflow. A researcher interested in epidemic rumours, for example, can filter the corpus for HEALTH_EPIDEMIC and compare stance patterns across periods and regions: are rumours mainly HEDGED, ATTRIBUTED, or DENIED_CORRECTED? A similar workflow applies to MARKETS_COMMERCE during financial panics, where correction labels may point to editorial interventions aimed at calming

markets or managing reputational risk. By linking each label to quoted evidence in the local span, the resource is designed to support exploratory mapping while keeping close reading central.

13. References

Allen, B., Sieczkiewicz, R., & Radev, D. (2020). What's hard about historical newspaper analysis? Technical Report, University of Michigan.

Bloch, M. (1921). Réflexions d'un historien sur les fausses nouvelles de la guerre. Revue de Synthèse Historique, 33, 13–35.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 964–971.

Mueller, M. (2014). Shakespeare His Contemporaries: Collaborative curation and exploration of early modern drama in a digital environment. Digital Humanities Quarterly, 8(3).

PleIAs. (n.d.). US-PD-Newspapers (Hugging Face dataset). https://huggingface.co/datasets/PleIAs/US-PD-Newspapers

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP 2019, 3982–3992.

Smith, D. A., & Cordell, R. (2018). A research agenda for historical and multilingual optical character recognition. NULab Working Paper.

biglam. (n.d.). hmd_newspapers (Hugging Face dataset). https://huggingface.co/datasets/biglam/hmd_newspapers

Author Index