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

Naomi Baes[Ψ] **Nick Haslam**[Ψ]
[Ψ] Melbourne School of Psychological Sciences
The University of Melbourne
n.baes,nhaslam@unimelb.edu.au

Abstract

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

Code: threshold-calibrated_wsd

1 Introduction

Polysemous words pose a methodological challenge for computational approaches to language change. Lexical semantic change is often studied using distributional methods that quantify semantic drift from contextualized embeddings, but interpreting these signals can require *sense-aware* modeling that distinguishes between a word's different contextual usages. For polysemous terms, distributional LSCD metrics are difficult to interpret because they do not directly encode sense distinctions and historical sense annotation is scarce. Therefore, it is often unclear to what extent observed change scores reflect shifts in underlying word senses.

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

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

The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26), pages 50–74

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

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

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

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

2 Related Work

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

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

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

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

3 Method

3.1 Corpus

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

3.2 Measures

3.2.1 Senses

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

3.2.2 Contextualized Embeddings

Each sentence was encoded using XL-LEXEME (XLL; Cassotti et al., 2023)[1], yielding a single 1024-dimensional *target-conditioned* embedding per sentence. We obtain this usage embedding by providing XLL with the target token span (start/end indices) for each sentence, producing a pooled representation explicitly conditioned on the marked target rather than a generic sentence embedding. XLL is a bi-encoder with a Siamese (SBERT-style) architecture (Reimers and Gurevych, 2019) fine-tuned on the Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2019) using contrastive learning to increase cosine similarity for sentence pairs expressing the same sense of a target word. This study specifies the target span via start-end indices rather than delimiter tokens, yielding a pooled vector sensitive to the marked target. We adopted XLL because it has demonstrated high sensitivity to induced semantic breadth among sentence-encoding baselines (Baes et al., 2025) and strong WiC performance relative to larger models (Periti and Tahmasebi, 2024). These vectors serve as word-in-context representations of the target term.

3.2.3 Graded Distributional Semantic Change

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

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

3.3 Threshold-Calibrated Sense Tracking Pipeline

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

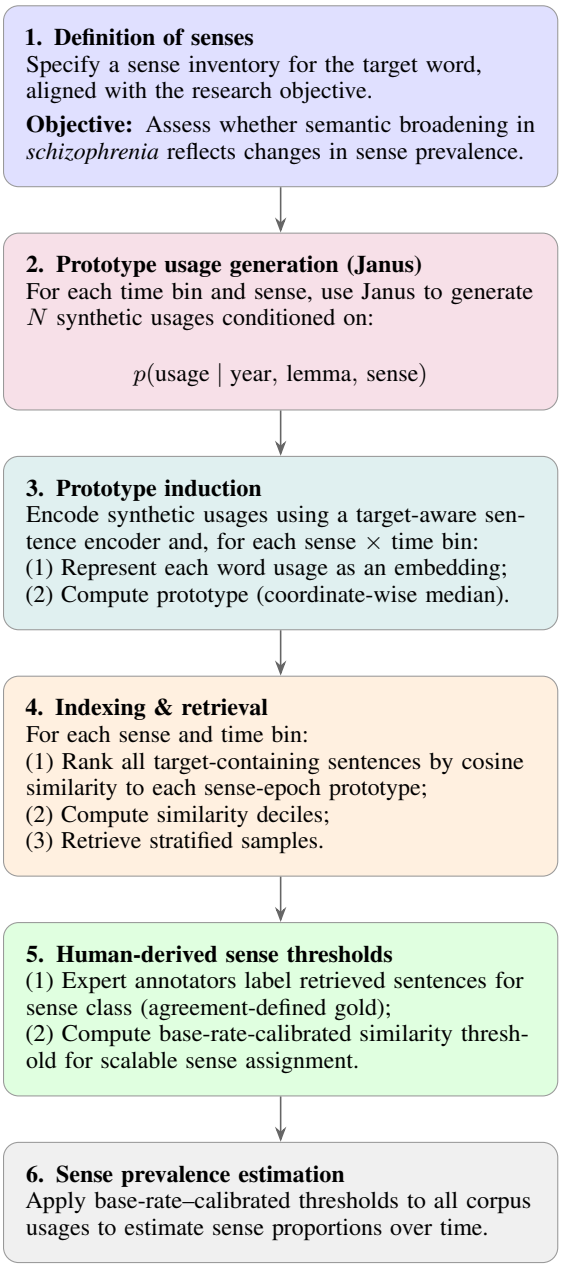

Figure 1: Six-stage threshold-calibrated sense tracking pipeline for hypothesis-driven, prototype-based sense assignment and prevalence tracking in text corpora.

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

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

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

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

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

[2]Model: ChangeIsKey/llama3-janus (8B parameters). https://huggingface.co/ChangeIsKey/llama3-janus.

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

[4]Duplicates due to generating 500 sentences per epoch.

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

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

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

(5) Human-derived sense thresholds. To enable scalable yet conservative sense assignment, we calibrated cosine similarity thresholds from expert annotations. For each sense, two annotators applied binary judgments (expresses the target sense or does not) using OED definitions; only unanimous labels were treated as gold. Because minority senses are rare, we used a two-round, stratified design to obtain reliable thresholds under severe class imbalance. In **Round 1**, we sampled uniformly across cosine similarity deciles for each sense prototype and computed *sense purity* (the proportion of unanimously labeled sentences matching the target sense) as a diagnostic check of whether a simple decile-based cutoff was possible. This strategy worked for the dominant psychiatric sense but yielded too few positives for the split-personality and metaphorical senses (which were underrepresented in the lower similarity deciles). We therefore conducted **Round 2**, enriching annotations in the highest-similarity region (top decile), where true positives for rare senses concentrate. Combining both rounds, we estimated each sense's corpus *base rate* by weighting similarity regions by their corpus mass, and set one global threshold per sense by selecting the cosine cutoff on the scored candidate set such that the number of retained sentences matched this estimated base rate, yielding conservative thresholds that limit over-assignment under class imbalance. Base-rate estimates and resulting thresholds are reported in Table 3; full estimator details and validation are provided in Appendix E.

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

4 Results

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

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

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

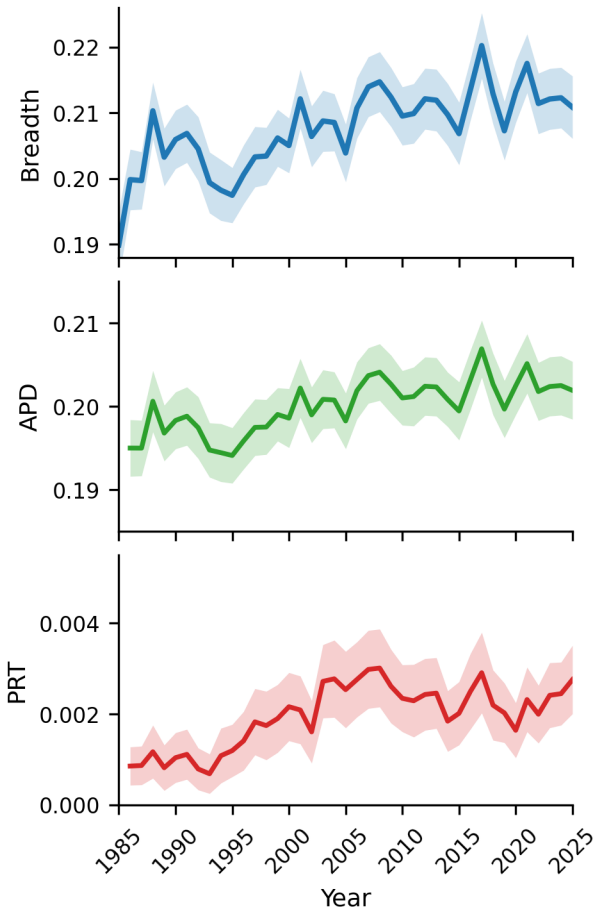

Figure 2: Lexical semantic change indices for *schizophrenia* in U.S. news (1985–2025). (A) Breadth: within-year contextual dispersion, reflecting heterogeneity in usage contexts. (B) APD: average pairwise cosine distance between usages in each year and those in the 1985 reference year, capturing baseline-relative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

Consistent with these patterns, time-series regression analyses summarized in Table 1 show significant positive temporal trends for Breadth and APD, indicating increasing contextual dispersion and cumulative baseline-relative distributional drift in uses of *schizophrenia*. These trends remain significant when controlling for annual sentence frequency, indicating that the observed changes are not reducible to increases in term prevalence. Results for Breadth remain significant under frequency-capping (500 sentences per year; $\beta = .00035$, CI = [.00024, .00052], $p < .001$, $R^2 = .55$), further supporting a robust increase in contextual semantic breadth. In contrast, although PRT demonstrates a positive temporal trend in the uncontrolled model, this effect attenuates and becomes non-significant once frequency is controlled.

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

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

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

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

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

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

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

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

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

To aid interpretation, we examined exemplar sentences selected using sentence-level thresholding (Appendix G). For each sense, we inspected both borderline-positive and high-confidence exemplars, illustrating the range of usages admitted by each category. Exemplars for Sense 1 (psychiatric) are qualitatively coherent across confidence levels, whereas exemplars for Senses 2 and 3b more frequently overlap with general or psychiatric mental-health contexts, indicating partial representational overlap. Consequently, small fluctuations in the estimated prevalence of non-dominant senses should be interpreted cautiously. A targeted follow-up analysis directly examined this dissociation and indicates increasing contextual heterogeneity within the dominant psychiatric sense (Appendix H).

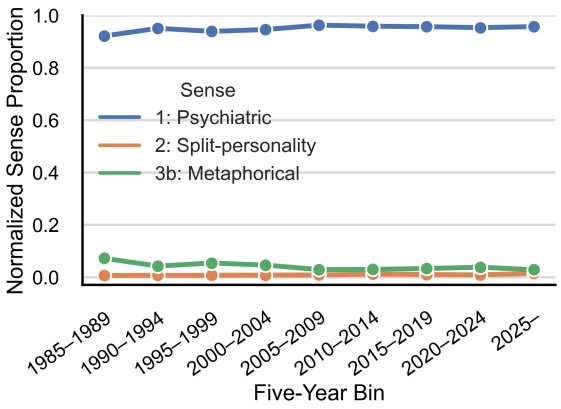

Figure 3: Relative prevalence of *schizophrenia* senses.

5 Discussion

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

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

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

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

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

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

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

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

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

6 Limitations

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

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

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

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

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

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

Finally, our analysis focuses on a single term in a single national news corpus, which limits generalizability and may reflect genre-specific editorial norms rather than broader patterns of semantic change across discourse domains (e.g., podcasts, scientific writing, or everyday language). The observed pattern — stable sense proportions alongside increasing contextual dispersion (particularly in schizophrenia's clinical sense) — is compatible with *determinologization* (Gorokhova, 2020), the linguistic process whereby specialized, technical, or scientific terms diffuse into broader public discourse thereby losing their strict, context-independent definitions to become part of the general vocabulary, while retaining a stable definitional core. Because the present study focuses on distributional and sense-prevalence dynamics rather than socioterminological boundary shifts, we do not model determinologization directly. Future work should apply this framework to additional polysemous terms (e.g., *intelligence*, *light*, *technology*), other harm-related concepts with multiple dictionary senses (e.g., *trauma*, *depression*), and less regulated discourse domains (e.g., social media), where metaphorical extension and terminological drift may be more prevalent and editorial constraints weaker. Words with more obviously disambiguated senses (e.g., *bank*) may show greater success using this threshold-calibrated word sense tracking method. Integrating formal determinologization frameworks with sense-aware dispersion diagnostics remains an important direction for future research.

Acknowledgments

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

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

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

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

part by Australian Research Council Discovery Project DP250102690.

References

Taichi Aida and Danushka Bollegala. 2025. Investigating the contextualised word embedding dimensions specified for contextual and temporal semantic changes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1413–1437, Abu Dhabi, UAE. Association for Computational Linguistics.

K. Bademli, A. Kaya Kılıç, and M. Kayakuş. 2023. Using twitter to assess stigma to schizophrenia and psychosis: A qualitative study. *Turkish Journal of Psychiatry*, 34(3):154–161.

Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2024. A multidimensional framework for evaluating lexical semantic change with social science applications. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1390–1415, Bangkok, Thailand. Association for Computational Linguistics.

Naomi Baes, Raphael Merx, Nick Haslam, Ekaterina Vylomova, and Haim Dubossarsky. 2025. LSC-eval: A general framework to evaluate methods for assessing dimensions of lexical semantic change using LLM-generated synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10905–10939, Vienna, Austria. Association for Computational Linguistics.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761, Online. Association for Computational Linguistics.

Joan L. Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press.

Belinda Cain, Roseanne Currie, Eleanor Danks, Fiona Du, Erica Hodgson, Jennifer May, Kirsty O'Loghlen, Yen Phan, Jennifer Powter, Nayab Rizwan, Shazmi Shahim, Dominique Simsion, Steve Loughnan, and Nick Haslam. 2014. "schizophrenia" in the australian print and online news media. *Psychosis*, 6(2):97–106.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Pierluigi Cassotti and Nina Tahmasebi. 2025a. A hypothesis-driven framework for detecting lexical semantic change. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 177–185, Cagliari, Italy. CEUR Workshop Proceedings.

Pierluigi Cassotti and Nina Tahmasebi. 2025b. Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics*, 13:690–708.

E. Castaño. 2023. What is in a word? an exploration of the metaphorical use of schizophrenia in general american english. *Lingua*, 294:103596.

Anju K. Chopra and Gillian A. Doody. 2007. Schizophrenia, an illness and a metaphor: Analysis of the use of the term "schizophrenia" in UK national newspapers. *Journal of the Royal Society of Medicine*, 100(9):423–426.

Sarah Delanys, Farah Benamara, Véronique Moriceau, François Olivier, and Josiane Mothe. 2022. Psychiatry on twitter: Content analysis of the use of psychiatric terms in french. *JMIR Formative Research*, 6(2):e18539.

Kenneth Duckworth, John H Halpern, Russell K Schutt, and Christopher Gillespie. 2003. Use of schizophrenia as a metaphor in us newspapers. *Psychiatric services*, 54(10):1402–1404.

Fabian Fabiano and Nick Haslam. 2020. Diagnostic inflation in the dsm: A meta-analysis of changes in the stringency of psychiatric diagnosis from dsm-iii to dsm-5. *Clinical Psychology Review*, 80:101889.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Natalia V Gorokhova. 2020. Determinologization and transterminologization processes in modern oil and gas discourse. In *European Proceedings of Social and Behavioural Sciences EpSBS*, pages 329–335.

Roksana Goworek, Harpal Singh Karlcut, Hamza Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Satkirti Purighella, Paridhi Gupta, Muhinyia Ndegwa, Bao Khanh Tran, and Haim Dubossarsky. 2025. SenWiCh: Sense-annotation of low-resource languages for WiC using hybrid methods. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 61–74, Vienna, Austria. Association for Computational Linguistics.

Nick Haslam. 2016. Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, 27(1):1–17.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Computational Approaches to Semantic Change*, pages 341–372. Language Science Press, Berlin.

Andrew J. Joseph, Neeraj Tandon, Lawrence H. Yang, Kenneth Duckworth, John Torous, Larry J. Seidman, and Matcheri S. Keshavan. 2015. #schizophrenia: Use and misuse on twitter. *Schizophrenia Research*, 165(2–3):111–115.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. *Northern European Journal of Language Technology*, 8.

Shalom Lappin. 2024. Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information*, 33(1):9–20.

Lorenza Magliano, John Read, and Riccardo Marassi. 2011. Metaphoric and non-metaphoric use of the term "schizophrenia" in Italian newspapers. *Social Psychiatry and Psychiatric Epidemiology*, 46(10):1019–1025.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Merriam-Webster. 2025. Schizophrenia. Merriam-Webster.com Dictionary.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69.

Oxford English Dictionary. 2025. Oxford english dictionary. Accessed 13 December 2025.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

ProQuest Dialog. 2013. How do i remove duplicate records? `https://pq-static-content.proquest.com/collateral/media2/documents/pqd-hdi-remove-duplicate-records.pdf`. ProQuest Support Center.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Dominik Schlechtweg, Frank D Zamora-Reina, Felipe Bravo-Marquez, and Nikolay Arefyev. 2025. Sense through time: Diachronic word sense annotations for word sense induction and lexical semantic change detection. *Language Resources and Evaluation*, 59(2):1431–1465.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, pages 1–91. Language Science Press.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

TDM Studio. 2023. Tdm studio. `https://tdmstudio.proquest.com`. ProQuest, part of Clarivate. Ann Arbor, MI, USA. Accessed: 2023–2025.

Elizabeth Closs Traugott and Richard B. Dasher. 2002. *Regularity in Semantic Change*. Cambridge University Press, Cambridge.

Sachin Yadav and Dominik Schlechtweg. 2025. Xl-durel: Finetuning sentence transformers for ordinal word-in-context classification. *arXiv preprint arXiv:2507.14578*.

A Corpus Construction and Preprocessing Details

A.1 Data Source

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

A.2 Deduplication

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

A.3 Sentence Extraction

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

A.4 Final Corpus

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

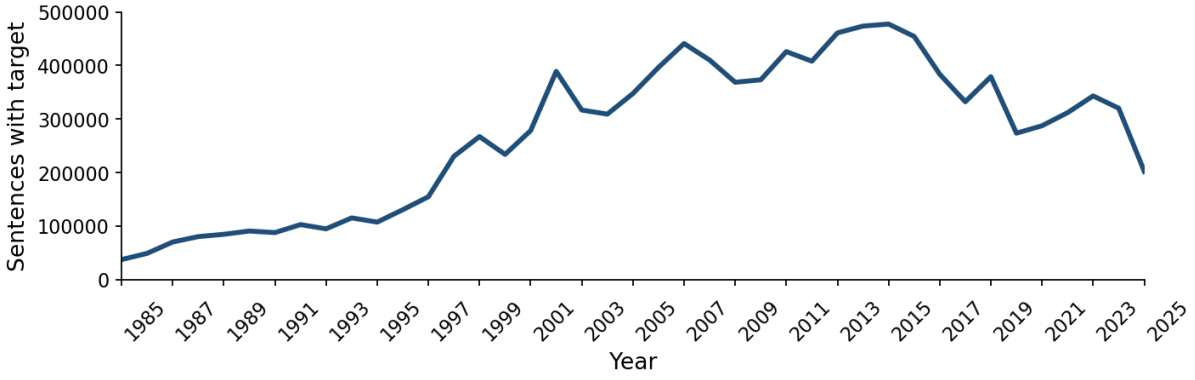

Figure 4: Count of sentences mentioning schizophrenia in the U.S. Newsstream corpus.

B Oxford English Dictionary Senses of Schizophrenia

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

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

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

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

C Graded Semantic Change Detection Metrics

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

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

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

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

Breadth Score (within-year).

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

Average Pairwise Distance (APD).

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

Prototype Representation Technique (PRT).

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

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

D Evaluation of Janus-Generated Diachronic Sense Prototypes

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

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

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

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

D.1 Illustrative Synthetic Usages

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

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

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

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

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

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

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

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

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

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

D.2 Quality Metrics of Sense Prototypes

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

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

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

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

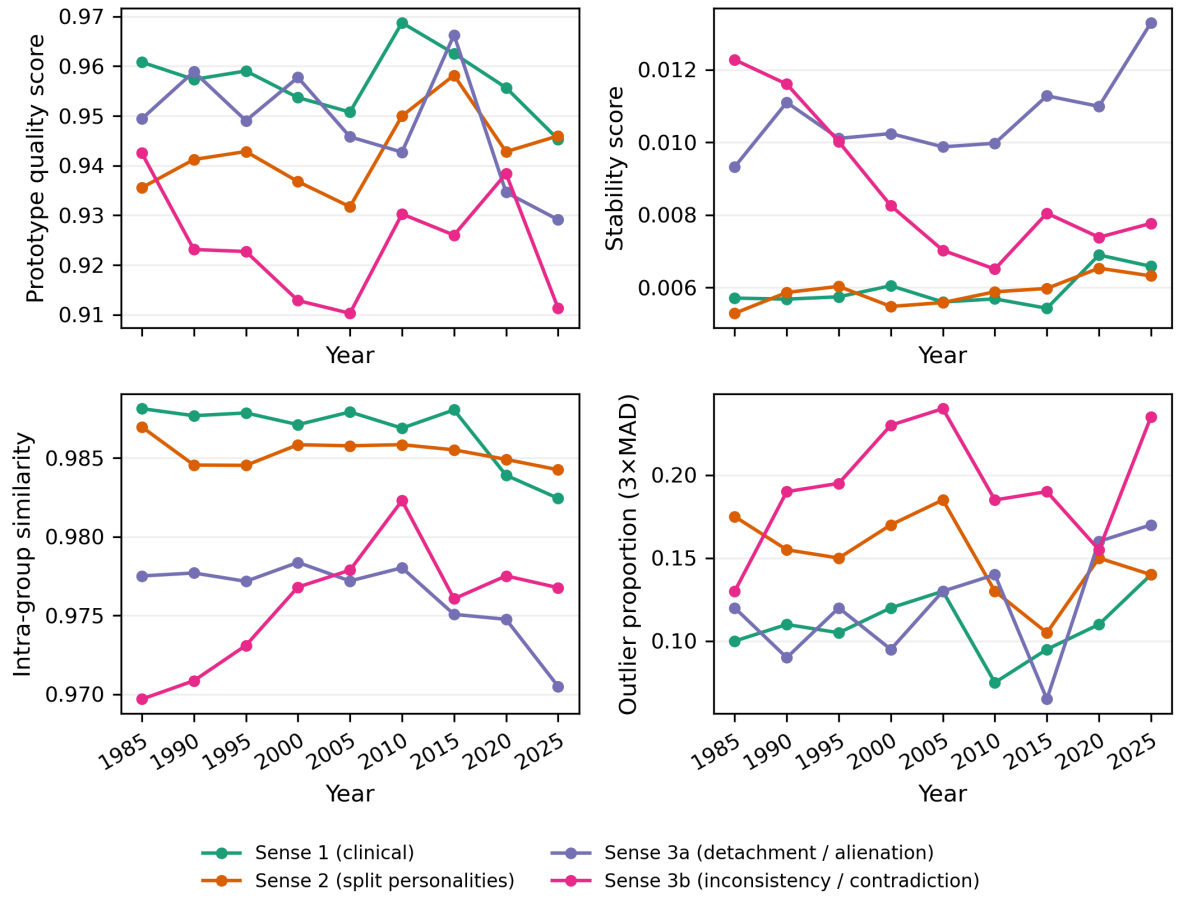

Figure 5: Local prototype-quality diagnostics for XL-LEXEME centroid embeddings across senses and years. (A) Prototype quality; (B) temporal stability; (C) intra-group similarity; and (D) outlier proportion ($3 \times$ MAD). Higher quality and intra-group similarity reflect more compact and coherent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

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

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

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

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

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

D.3 Descriptive Statistics for Prototype–Sentence Cosine Similarity

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

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

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

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

E Annotation Diagnostics and Threshold Estimation

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

E.1 Human Annotation Protocol

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

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

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

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

E.2 Annotation Diagnostics and Threshold Estimation

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

E.2.1 Round 1: Stratified annotation diagnostics

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

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

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

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

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

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

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

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

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

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

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

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

F Sense Proportions Diagnostics

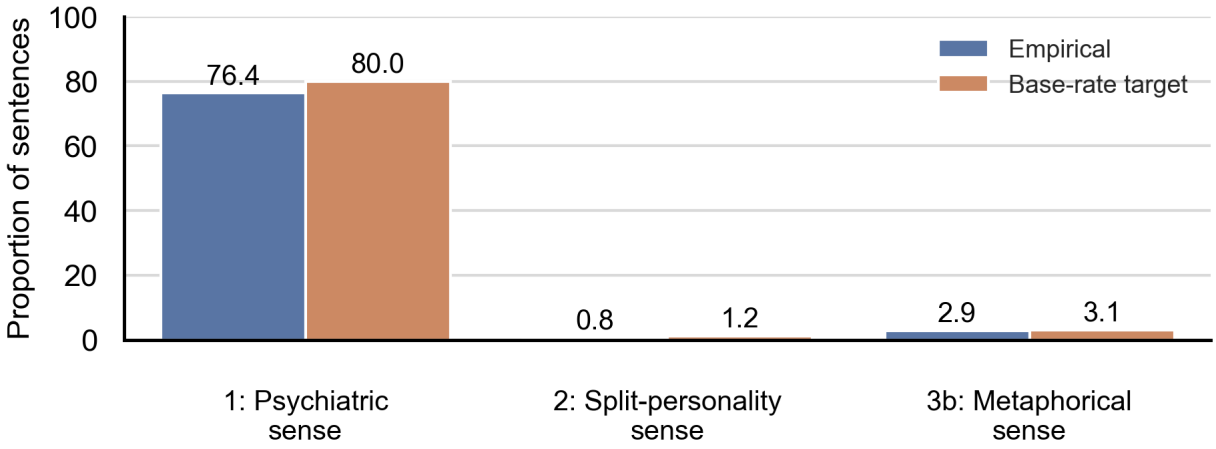

Figure 6: Global proportions versus base-rate targets.

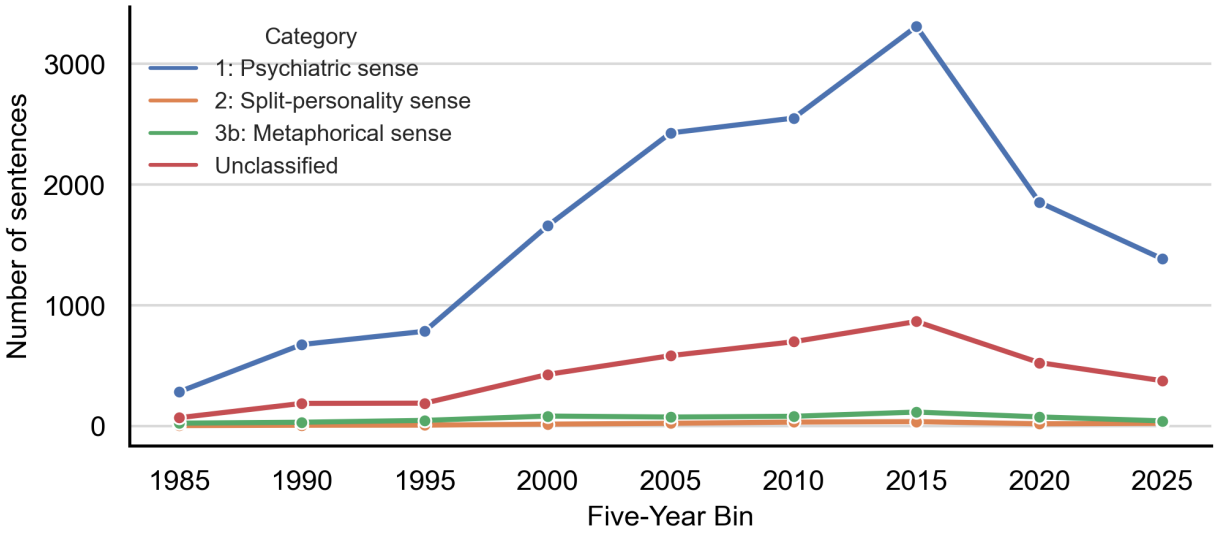

Figure 7: Sense counts over time (assigned senses and unclassified).

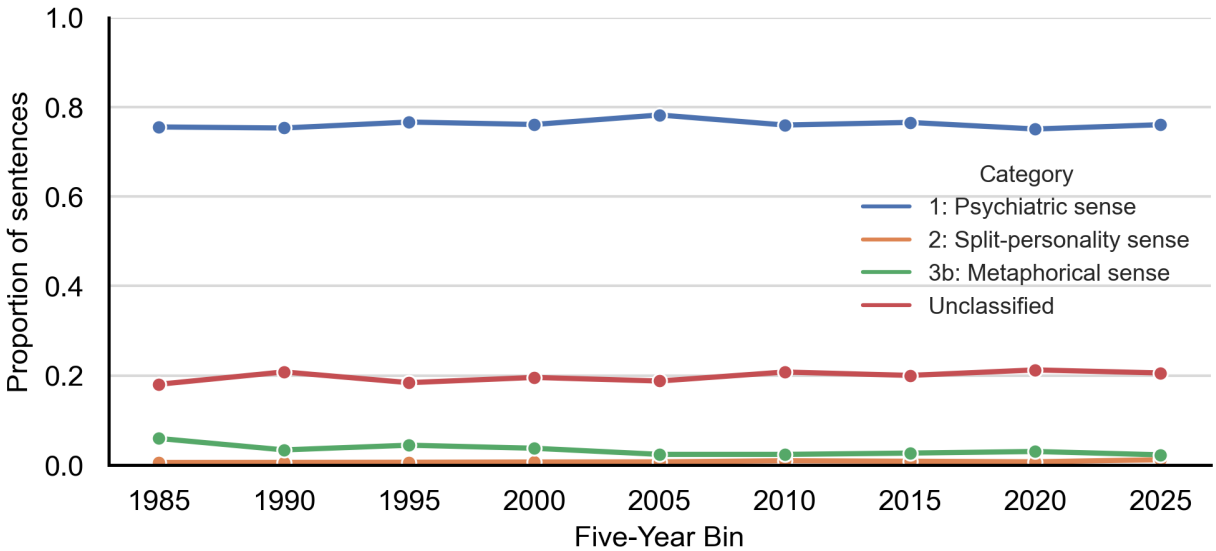

Figure 8: Sense proportions over time (base-rate-calibrated, including unclassified).

G Threshold exemplars for interpreting sense proportions

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

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

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

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

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

H Intra-sense heterogeneity within the psychiatric sense

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

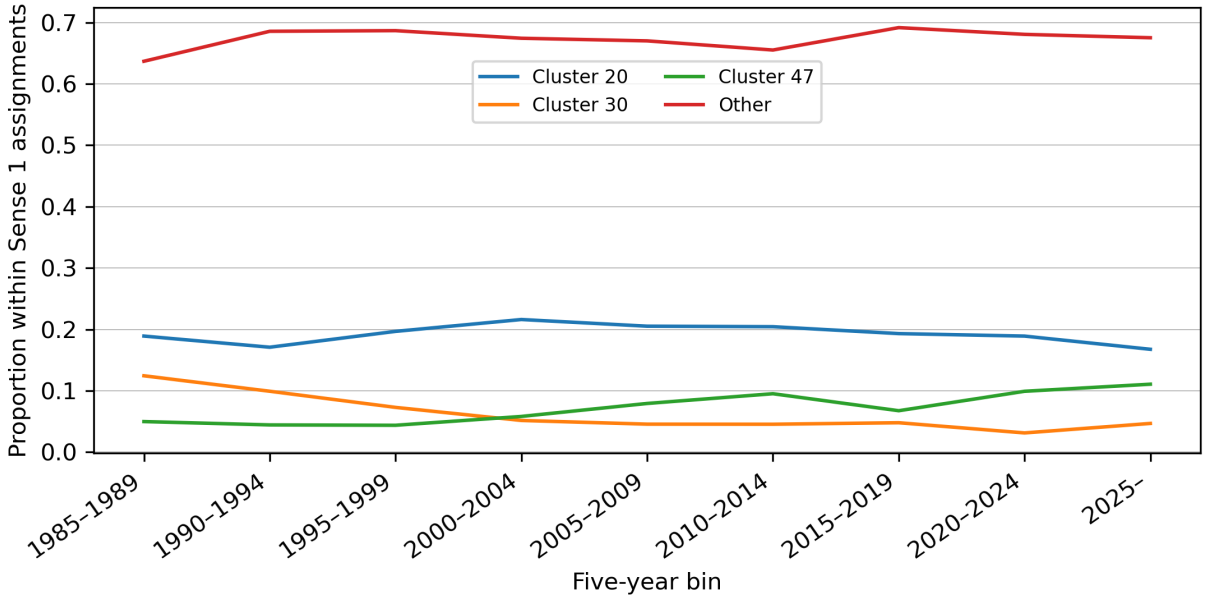

Figure 9: Relative prevalence of the three largest HDBSCAN subclusters within Sense 1 across five-year bins. Proportions are normalized within Sense 1 assignments per bin; remaining clusters are grouped as *Other*.

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

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

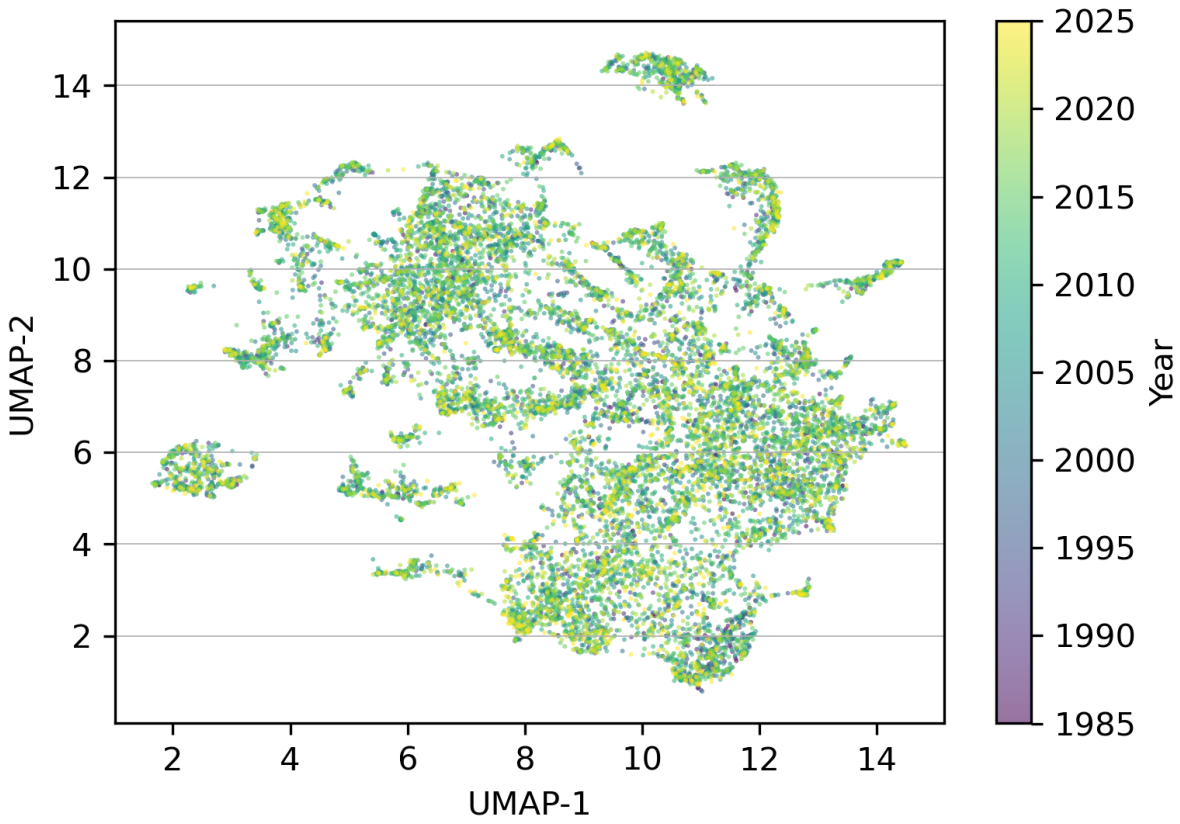

Figure 10: UMAP projection of Sense 1-assigned usages colored by year. Points from different decades largely overlap, indicating temporal continuity within the psychiatric sense.

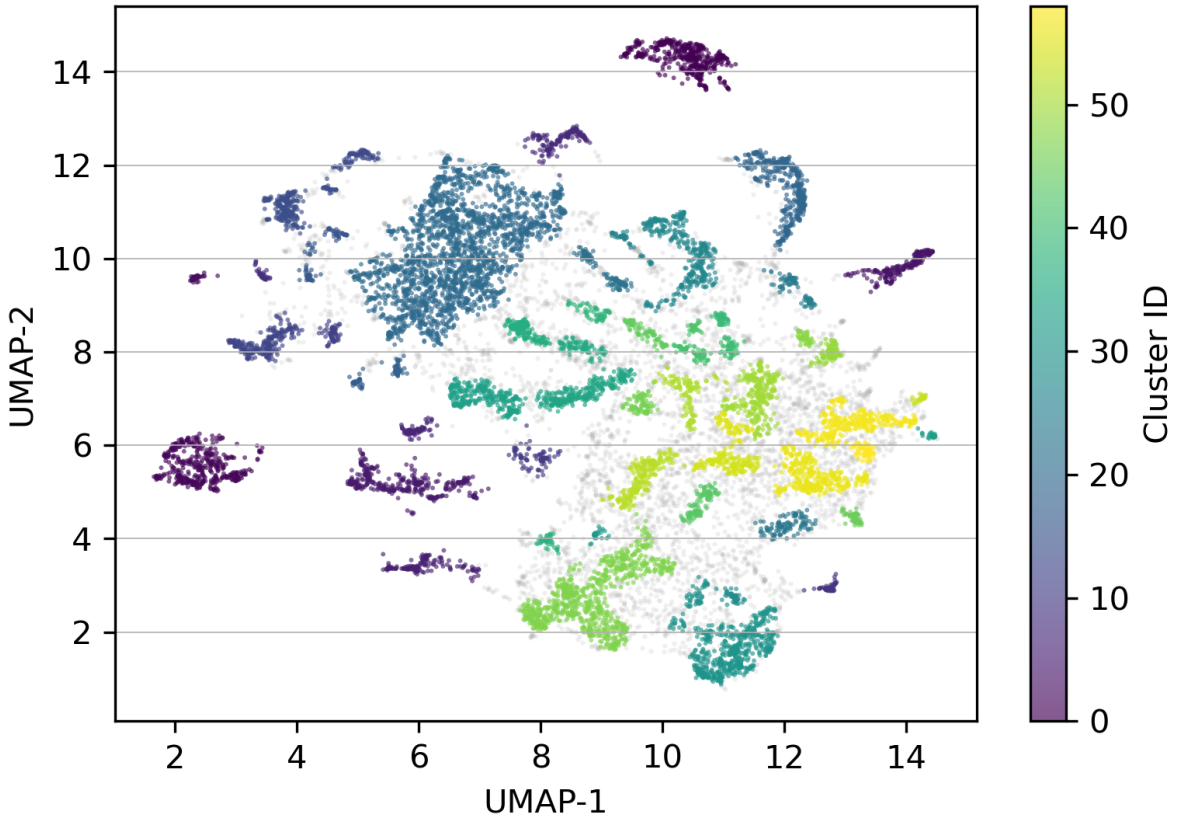

Figure 11: UMAP projection of Sense 1-assigned usages colored by HDBSCAN cluster. Gray points indicate noise assignments.