A Multidimensional Framework for Evaluating Lexical Semantic Change with Social Science Applications

Naomi Baes[Ψ] **Nick Haslam**[Ψ] **Ekaterina Vylomova**[λ]
[Ψ] Melbourne School of Psychological Sciences
[λ]School of Computing and Information Systems
The University of Melbourne
{n.baes, nhaslam, vylomovae}@unimelb.edu.au

Abstract

Historical linguists have identified multiple forms of lexical semantic change. We present a three-dimensional framework for integrating these forms and a unified computational methodology for evaluating them concurrently. The dimensions represent increases or decreases in semantic 1) sentiment (valence of a target word's collocates), 2) breadth (diversity of contexts in which the target word appears), and 3) intensity (emotional arousal of collocates or the frequency of intensifiers). These dimensions can be complemented by the evaluation of shifts in the frequency of the target words and the thematic content of its collocates. This framework enables lexical semantic change to be mapped economically and systematically and has applications in computational social science. We present an illustrative analysis of semantic shifts in *mental health* and *mental illness* in two corpora, demonstrating patterns of semantic change that illuminate contemporary concerns about pathologization, stigma, and concept creep.

1. Introduction

Lexical semantic change is defined by historical linguists as innovations that alter the meaning, but not the grammatical function, of a form (Campbell, 1999). For instance, "awesome" once denoted the capacity to inspire awe, but its meaning has since been bleached to a general expression of approval. Computational linguists have made strides in developing distributional semantic methods (Boleda, 2020) to detect semantic change (Kutuzov et al., 2018; Tahmasebi et al., 2021; Tang, 2018) and its laws (Hamilton et al., 2016b) as distinct from cultural shifts (Hamilton et al., 2016a).

Advances in deep learning since 2018 (Manning, 2022) afford new ways to model semantic change processes. These innovations have facilitated the development of language models with sophisticated word embeddings or vector representations. As a result, word embeddings have evolved from count-based models (Jurafsky and Martin, 2023), where words are represented by their co-occurrence frequency with other words, to prediction-based representations (Mikolov et al., 2013; Pennington et al., 2014), where word vectors are iteratively learned as part of a language modelling task objective. The granularity of these representations shifted from *type-level*, where each word has a single vector despite its usages, to *token-based*, or contextualized representations (Montanelli and Periti, 2023; Kutuzov et al., 2022), where each word instance (token) has a vector, dynamically capturing shifts in meaning based on context. Lexical semantic relations can be detected by type- (Shwartz et al., 2016; Vylomova et al., 2016) and token-level (Rogers et al., 2020) embeddings.

Other work has started addressing the challenge of formalizing and understanding kinds of semantic change (Hengchen et al., 2021). Processes such as broadening (Vylomova et al., 2019; Yüksel et al., 2021), metaphorization (Maudslay and Teufel, 2022), and pejoration/amelioration (Fonteyn and Manjavacas, 2021) have been modelled. Researchers have created methods to automatically disambiguate a word's pejorative usage from its non-pejorative use (Dinu et al., 2021). Attempts have also been made to evaluate understudied classes of semantic change. Sentence representations from neural language models were used for hyperbole detection (Schneidermann et al., 2023). Exaggerated language can be generated (Tian et al., 2021) and detected (Kong et al., 2020), alongside metaphor (Badathala et al., 2023). Researchers have also evaluated semantic bleaching, whereby words lose elements of their meaning (Luo et al., 2019), and found it to be triggered in contexts where an adverb premodifies a semantically similar adjective (e.g., "insanely jealous"). Nevertheless, there are a dearth of diachronic methods for evaluating lexical semantic change (de Sá et al., 2024).

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1390–1415

Despite advances in detecting and modelling lexical semantic change, there is a need for a unifying framework to integrate multiple dimensions of change. The present study addresses this gap by proposing a framework which synthesizes the theoretical insights of historical linguists about the many distinct forms of diachronic lexical semantic change (e.g., Bloomfield, 1933) and aligns them with the methodological sophistication of natural language processing. The comprehensive computational framework for evaluating lexical semantic change that emerges should be valuable for computational social scientists seeking to understand and model social and cultural change.

2. Related Work

2.1 Forms of Lexical Semantic Change

Historical linguists have developed several taxonomies of the forms of lexical semantic change (Blank, 1999; Bréal, 1897; Ullmann, 1962), but Bloomfield's (1933) is one of the most well-established. Bloomfield described nine forms identified by earlier scholars: (1) narrowing: superordinate to subordinate, or when a meaning becomes more restricted (Old English *mete* 'all food' > *meat* 'edible flesh'); (2) widening: subordinate to superordinate, or specific to general expansion of meaning (Middle English *dogge* 'dog of a specific breed' > dog); (3) metaphor: the transfer of a name based on the associations of similarity or hidden comparison (Primitive Germanic *bitraz* 'biting', derivative of 'I bite' > *bitter* 'harsh of taste'), (4) metonymy: change based on the meanings' proximity in space or time (Old English *ceace* 'jaw' > *cheek*); (5) synecdoche: the meanings are related as whole and part (pre-English *stobo* 'heated room' > stove), (6) hyperbole: stronger to weaker meaning by overstatement (pre-French *extonare* 'to strike with thunder' > to astonish; English borrowed *astound*, *astonish* from Old French); (7) meiosis:[1] weaker to stronger meaning by understatement (pre-English *kwalljan* 'to torment' > Old English *cwellan* 'to kill'); (8) degeneration: positive to negative connotation (Old English *cnafa* 'boy servant' > *knave*); (9) elevation: negative to positive connotation (Old English *cniht* 'boy, servant' > *knight*).

Bloomfield's classes align closely with the forms of change identified in studies of denotational and connotational meaning (Geeraerts, 2010). For denotational (referential) meaning, Geeraerts identifies (1) specialization, (2) generalization, (3) metonymy, and (4) metaphor. Specialization (semantic 'restriction' and 'narrowing') implies that the new meaning covers a subset of the old meaning's range; for generalization (or 'expansion', 'extension', 'schematization', 'broadening'), the new range includes the old meaning. Metonymy (here including synecdoche) is a "link between two readings of a lexical item based on a relationship of contiguity between the referents of the expression in each of those readings" (Geeraerts, 2010, p. 27). Conversely, metaphor is based on similarity. Geeraerts also identifies two forms of connotational meaning (i.e., the aspects of a word's meaning that are related to the writer or reader's emotions, sentiment, opinions, or evaluations): (1) pejorative and (2) ameliorative change (i.e., shift towards a more negative/positive emotive meaning). An example of pejoration is 'silly', which formerly meant 'deserving sympathy, helpless', but has come to mean 'showing a lack of common sense'. Amelioration is shown by 'knight' once meaning 'boy, servant'.

2.2 Expanding Concepts of Harm and Pathology

Semantic change processes such as these may partly reflect cultural, social, and political shifts, and are of interest to social science researchers. One example is social psychological research on concept creep, the semantic expansion of harm-related concepts (e.g., abuse, bullying, mental illness, prejudice, trauma, violence; Haslam, 2016). Concept creep takes two forms: harm-related concepts have expanded 'horizontally' to cover a wider range of harms and 'vertically' to encompass less intense harms. It is theorized to be driven by rising cultural sensitivity to harm (Furedi, 2016; Wheeler et al., 2019), falling societal prevalence of harm (Levari et al., 2018; Pinker, 2011), and deliberate conceptual expansion by "opprobrium entrepreneurs" (Sunstein, 2018). Concept creep is theorized to have mixed blessings (Haslam et al., 2020), trivializing harms on one hand (Dakin et al., 2023) and enhancing the recognition and redress of major harms on the other (Tse and Haslam, 2021).

Prior empirical work has evaluated concept creep in historical text corpora. Studies assessing horizontal expansion as increases in the broadening of harm concepts found that some concepts (e.g., addiction, bullying, trauma) have broadened within academic psychology (Haslam et al., 2021; Vylo-

[1] Bloomfield (1933) refers to this class as litotes, but we use meiosis to reflect general understatement.

mova et al., 2019; Vylomova and Haslam, 2021). Recent work evaluated the vertical form of concept creep, defined as the concept's use in contexts of declining emotional intensity, and yielded mixed findings for anxiety, depression, grief, stress, and trauma (Baes et al., 2023a,b; Xiao et al., 2023).

Mental illness has become an increasingly salient term in society (Haslam and Baes, 2024), partly due to the recent prioritization of mental health in global health policy (WHO, 2021). Critics have raised concerns that the rising prominence of mental health discourse is instigating problematic changes in how people conceptualize mental ill health. Some contend that concepts of mental illness have broadened so that everyday life is increasingly pathologized (Brinkmann, 2016; Horwitz and Wakefield, 2007, 2012). Experiences that were once considered normal are now given diagnostic labels, such as using 'depression' to reference ordinary sadness (Bröer and Besseling, 2017). Alternatively, it has been argued that terms like "mental health problems" are being normalized and broadened (Sartorius, 2007), alongside increasing prevalence of mental illnesses. Some argue that concepts of mental illness are becoming less stigmatizing, although this question has only been addressed in surveys of public attitudes (e.g., Schomerus et al., 2022), rather than in changes in word connotations. In view of the widespread speculation on the ways in which concepts of mental illness have changed historically and the lack of scientific evidence of these shifts, a systematic study of conceptual change in this domain is a priority.

2.3 Our Original Contribution

The present study aims to make three main contributions: (1) it proposes a multidimensional framework for evaluating lexical semantic change that economically integrates forms identified by historical linguists; (2) it develops a set of computational methodologies for evaluating change on these dimensions; and 3) it illustrates this computational framework by examining semantic shifts in concepts of mental health and mental illness to address cultural concerns about pathologization, normalization, and stigmatization. The study will therefore test if the framework can thoroughly illuminate how *mental health* and *mental illness* have changed their meanings in two corpora representing academic psychology and general US English text.[2]

[2]The source code is available here: https://github.com/naomibaes/lexical_semantic_change_framework

3. Method

3.1 Framework

The proposed framework, illustrated in Figure 1, economically reduces classes of lexical semantic change identified by historical linguists (excluding metaphor and metonymy; Geeraerts, 2010) to three dimensions. It recognizes that these classes represent opposed pairs of change types, each member corresponding to a pole on a single dimension. In essence, the framework reformulates six classes as three dimensions, allowing lexical semantic change to be quantified on three axes simultaneously rather than categorized into exclusive types. A recent survey paper (de Sá et al., 2024) has also classified semantic change as having three classes of characterizations related to a word's meaning becoming used in a more (1) pejorative or ameliorated sense (orientation), (2) metaphoric or metonymic context (relation), (3) abstract/general or more specific/narrow context (dimension). However, their theoretical framework does not consider hyperbole/litotes.

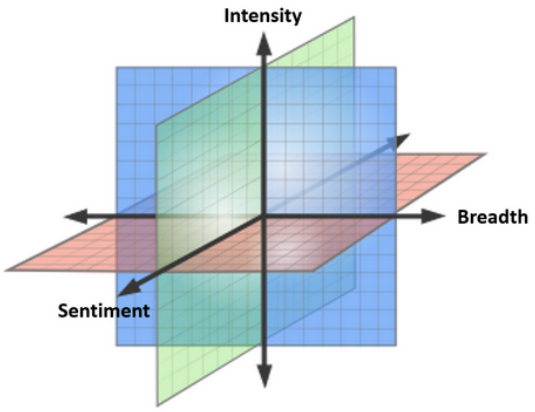


Figure 1: Three Major Dimensions of Semantic Change.

In our proposed framework, the *Sentiment* dimension relates to whether the word acquires a more positive ('elevation', 'amelioration') or negative ('degeneration', 'pejoration') connotation. The *Breadth* dimension relates to whether a word expands ('widening', 'generalization') or contracts ('narrowing', 'specialization') its semantic range. The *Intensity* dimension relates to whether a word changes to refer to more emotionally or referentially intense phenomena ('meiosis') or less intense phenomena ('hyperbole'). Table 1 summarizes how the three dimensions map onto the classes of lexical semantic change as well as the two proposed forms of concept creep.

Dimension	Rising	Falling
Sentiment	Elevation (Bloomfield, 1933); Amelioration (Ullmann, 1962)	Degeneration (Bloomfield, 1933); Pejoration (Ullmann, 1962)
Breadth	Widening (Bloomfield, 1933; Ullmann, 1962); Generalization of meaning (Blank, 1999); Horizontal Creep (Haslam, 2016)*	Narrowing (Bloomfield, 1933; Ullmann, 1962); Specialization of meaning (Blank, 1999)
Intensity	Meiosis (Bloomfield, 1933)	Hyperbole (Bloomfield, 1933); Vertical Creep (Haslam, 2016)*

Table 1: Dimensions of Lexical Semantic Change and their associated forms. * = specific to harm-related concepts.

The three proposed dimensions align with established dimensions in other domains. For example, Sentiment and Intensity resemble the two primary dimensions of human emotion, Valence and Arousal (Russell, 2003), and two primary dimensions of connotational meaning, Evaluation (e.g., "good/bad") and Potency (e.g., "strong/weak") (Osgood et al., 1975), both of which have been shown to have cross-cultural validity. Although our dimensions capture the primary forms of lexical change, we argue that they can be complemented by evaluation of changes in a word's salience (i.e., relative frequency of use) and its *thematic content* (i.e., shifts in the specific contexts in which the word is used). These dimensions may reflect psychological, sociocultural, or cultural forces that contribute to or result from semantic change (Blank, 1999). Our case study of *mental health* and *mental illness* illustrates how attention to salience and thematic content enrich the characterization of semantic change that the three primary dimensions provide. We now turn to the details of that case study, including the computational methodologies for evaluating these dimensions. Future implementations of our three-dimensional framework are likely to include technical refinements of these methodologies. Those employed in the case study simply demonstrate one way to implement it using interpretable techniques.

3.2 Sentiment

The sentiment of the target concepts (*mental health* and *mental illness* and the control concept *perception*) was evaluated using valence norms from Warriner et al. (2013), which provide valence ratings for 13,915 English lemmas collected from 1,827 United States residents, ranging from low valence (1: feeling extremely "unhappy", "despaired") to high valence (9: feeling extremely "happy", "hopeful"). See Appendix A for more information regarding the valence ratings. Collocates of each target concept were extracted within a ± 5-word context window (Agirre et al., 2009) and matched to the Warriner et al. norms which showed adequate coverage for the psychology corpus but poorer coverage for the general corpus ("mental_health": psychology = 84%; general = 50%; "mental_illness": psychology = 83%;general = 48%; "perception": psychology = 84%; general = 39%). Annual counts of Warriner-matched collocates for each target concept were then extracted from the lemmatized corpora, which showed few occurrences due to few appearances of texts containing targets before 1990 in the general corpus (see Appendix B). Therefore, analyses excluded general texts before 1990. The annual sentiment score for each concept was computed by weighting the valence rating for each collocate by its annual appearances, standardized by the total number of (matched) collocates in the respective year. The index represents the mean valence of terms [1,9] collocating with target concepts, where higher scores indicate higher valence.

3.3 Breadth

The semantic broadening of the target concept was evaluated as the average inverse cosine similarity between the sentence level embeddings containing the target term. Our method adapts previous work (Vylomova et al., 2019; Vylomova and Haslam, 2021) by replacing type-level word embeddings with contextualized sentence-level embeddings. Given that this breadth measure resembles the Semantic Textual Similarity (STS) task (Cer et al., 2017, the degree to which two sentences are semantically equivalent to each other), to select the optimal model we compared the sentence similarity scores, from corpus samples, of models that have shown good performance for encoding sentences. Many of the original Sentence-BERT models (Reimers and Gurevych, 2019) with good scores on semantic textual similarity benchmarks

(Tsukagoshi et al., 2022; Reimers and Gurevych, 2019) are deprecated, therefore we examined and compared three public pre-trained models that currently excel in encoding sentences,[3] from the sentence transformers library. See Appendix C for more information regarding model selection (C), comparison (C) and results (C). The pre-trained model used in the present study[4] performed best on detecting *semantic* information and encoding sentences for 14 diverse tasks from different domains.

To compute the breadth score, relevant texts were extracted from our corpora. Inspecting their frequencies showed that it was acceptable to sample 50 texts from each five-year interval.[5] Thus, we randomly and uniformly sampled up to 50 sentences per interval and repeated the procedure 10 times to reduce sampling noise. These sentences were then passed to the sentence transformer model, "all-mpnet-base-v2" (where MPNET means Masked Permuted Language Modeling Network), to be tokenized and to encode embeddings representing their semantic characteristics. Cosine distance was computed for each pair of sentence vectors by inverting the similarity scores (1 - cosine similarity). The final breadth metric [0,1] was calculated by averaging scores across samples in each interval. Higher scores indicate greater breadth (*dis*similarity) between sentence vectors.

3.4 Intensity

Changes in the intensity of the concepts were evaluated in two ways. First, we computed an arousal index, adapting a previously established procedure (Baes et al., 2023a,b; Xiao et al., 2023). In an equivalent manner to the sentiment analysis, we examined the collocates of each concept and computed a weighted average annual ratings, using Warriner et al.'s arousal norms that range from low arousal (1: feeling "calm", "unaroused" while reading the lemma) to high arousal (9: feeling "agitated", "aroused"). See Appendix A for more information regarding arousal ratings. The annual arousal score for each concept was calculated by weighting the arousal rating for each collocate by its total number of appearances in each year and normalizing it by the total (matched) collocate count for the respective year. The index represents the mean arousal of terms [1,9] collocating with target concepts, where higher scores indicate higher arousal.

Second, we developed a new index to directly capture shifts in a concept's intensity. Instead of examining the arousal of its collocates (regardless of their order), it examined the occurrence of intensifying expressions that directly modify it. If a concept increasingly appears with an intensifying modifier, it can be inferred that its unmodified meaning has become less intense. We developed a new "intensifier index" which evaluates the relative frequency with which 11 adjectival modifiers ("great", "intense", "severe", "harsh", "major", "extreme", "powerful", "serious", "devastating", "destructive", "debilitating") preceded "mental health" and "mental illness". De-adjectival adverbs from Luo et al. (2019) were considered but most were not sufficiently general (e.g., "devastating", "excruciating", "vicarious"). We used the dependency-parsed corpora (see Section 4.2) to compute the proportion of instances of each target concept that has any of the 11 terms as its adjective modifier.

3.5 Thematic content

Thematic content was evaluated using a top-down approach. The theme of interest was pathology given concerns raised by critics about the pathologization of *mental health* and *mental illness* (Brinkmann, 2016; Horwitz and Wakefield, 2007, 2012). We used a pathologization dictionary developed by Baes et al. (2023a) to compute the pathologization index. This approach can be used to construct dictionaries for other themes of interest. First, we generated unambiguously disease-related words with restricted range in meaning: "clinical", "disorder", "symptom", "illness", "pathology", and "disease". Next, their forward word associations (participant responses to each disease-related word) drawn from the English Small World of Words project (De Deyne et al., 2019) were listed and duplicates were removed. We filtered the list for terms reflecting pathologization (i.e., to view or characterize as medically or psychologically abnormal), leaving 17 terms: "ailment", "clinical", "clinic", "cure", "diagnosis", "disease", "disorder", "ill", "illness", "medical", "medicine", "pathology", "prognosis", "sick", "sickness", "symptom", "treatment". Following Baes et al. (2023a), we computed the pathologization index by dividing appearances of the 17 terms in the target concept's collocates (±5-word context window) in a specific year by the total number of collocates in that year.

[3] https://www.sbert.net/docs/pretrained_models.html

[4] "all-mpnet-base-v2" from Hugging Face, sentence-transformers: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[5] Appendix C explains interval selection.

3.6 Salience

Salience was computed as the concept's annual relative frequency, using the raw corpora versions.

4. Materials

4.1 Corpora

Two corpora were chosen for their historical length, their magnitude, and their texts. The psychology corpus contained 143,575,773 tokens from 871,344 abstracts from 875 (Scimago indexed) psychology journals, ranging from 1930 to 2019, sourced from E-Research and PubMed databases (Vylomova et al., 2019). The journal set was distributed across all subdisciplines of psychology. The final corpus of psychology abstracts was limited to 1970-2016 due to the relatively small number of abstracts outside this period (Vylomova et al., 2019), yielding 129,980,596 tokens from 793,942 abstracts.

The second corpus is a combination of two related corpora: the Corpus of Historical American English (Davies, 2010, 1810-2009) and the Corpus of Contemporary American English (Davies, 2008, 1990-2019). Academic texts were excluded to avoid any potential overlap with psychology articles. After merging the two corpora, containing 115,000 everyday publications and >500,000 contemporary texts, the combined corpus was processed following recommendations from Alatrash et al. (2020) to maintain data integrity.[6] The current study restricted the corpus period from 1970 to 2016, using 501,415,577 tokens from 244,552 texts (books: 23,855 fiction, 1,498 non-fiction; 88,641 magazines; 73,557 newspapers; 40,036 spoken language; 16,965 TV shows).

4.2 Preprocessing

Analyses required three versions of the corpora: (1) a raw cleaned version transforming target concepts to single noun tokens (Section 3.6 and 3.3 and 3.4); (2) a lemmatized version (Section 3.2, 3.4, and 3.5); and (3) a dependency parsed version (Section 3.4). The first version, including punctuation, uppercasing, and numbers, was used for all analyses after transforming multiword target concepts into single tokens (e.g., "mental health" > "mental_health") using case sensitive matching. The lemmatization pipeline included tokenization, part-of-speech tagging (skipping tokens with uninformative tags: punctuation, symbols, spaces, numbers), removing stop words (uninformative words like "the"), and lemmatization using spaCy.[7] For dependency parsing we used the raw corpora to provide more contextual information for the model to better understand relationships between words. The English Transformer model[8] was used to preprocess the corpus with a high performance computing system (Lafayette et al., 2016).

4.3 Target Concepts

Two terms were chosen to analyze levels of semantic change (Hamilton et al., 2016a): *mental_health* and *mental_illness*. We also ran control analyses using the neutral term, *perception*, for which a fixed rate of change was expected and which demonstrated a steady rise in relative frequency starting around 1945 in the Google Ngram Viewer.[9]

4.4 Statistical Analysis

Linear regression analyses were performed to test the statistical significance of historical trends in the semantic indices (Jebb et al., 2015). Ordinary least squares served as the primary estimator, the secondary one being a generalized least squares estimator to account for auto-correlated residuals (Durbin-Watson test: $p < .05$). Coefficients, standard errors and confidence intervals were standardized using the betaSandwich package (Pesigan et al., 2023), employing Dudgeon's (2017) heteroskedasticity-consistent estimator approach (HC3), ideal for extracting estimates for nonnormal data and small sample sizes (Dudgeon, 2017). The code is publicly available.[10]

5. Results

Sentiment: The linear regression models mostly show decreasing trends for the valence index. Figure 2 shows a significant declining trend in the valence of words used in the context of *mental health* in the psychology corpus and the general corpus. For *mental illness*, the valence index shows a decreasing trend in psychology, and an increase in the general corpus. The valence of *perception* only shows a decreasing trend in the general corpus.

Breadth: The linear regression models testing the trend for the cosine distance of sentential contexts containing targets show significant increas-

[6] See Appendix D for a comprehensive explanation.

[7] https://spacy.io/

[8] "en_core_web_trf" (roberta-base) from Spacy was used as it demonstrates the highest accuracy on 13 evaluation tasks: https://spacy.io/models/en#en_core_web_trf.

[9] https://books.google.com/ngrams/info

[10] https://osf.io/4d7ur/

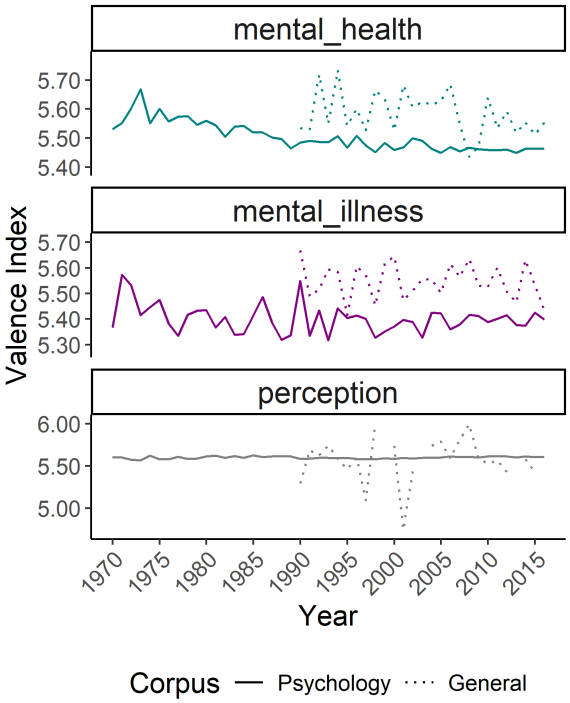


Figure 2: Valence index over the study period (1970-2016).

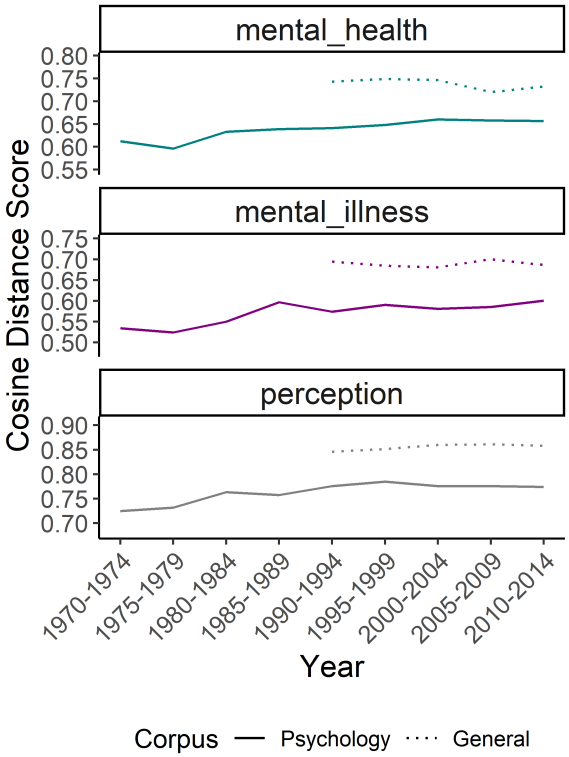


Figure 3: Breadth score over five-year intervals (1970-2014).

ing trends for *mental health*, *mental illness* and *perception* in the psychology corpus, reflecting greater sentence diversity, with a decrease for *mental health* and an increase for *perception* in the general corpus, as shown in Figure 3.

Intensity: Figure 4 shows the significant rise and fall in the use of intensifiers to modify *mental illness* in the psychology corpus, but no trend in the general corpus. Examining the top ranked adjective modifiers in each decade (Table 4 and Table 7 in Appendix E) reveals that "severe", "serious", "major", "chronic" come to be more associated with *mental illness* from the 1990s onwards. Although *mental health* is not frequently modified by intensifiers, as expected, "poor" and "positive" remain closely associated with it across the decades, with "maternal" becoming more associated with *mental health* from the 1990s onwards. Despite demonstrating a significant increase in its intensifier index in the psychology corpus, *perception* does not display intensifiers among its top adjective modifiers.

Figure 5 shows a significant increasing trend in the intensity (arousal index) of *mental health*-related words in both corpora. For *mental illness* and *perception*, the index increases significantly for the psychology corpus and only shows an increasing trend for *perception* in the general corpus.

Thematic content: The target concepts, *mental health* and *mental illness*, and the control *perception*, become significantly more associated with pathology-related terms in the psychology corpus, and for all targets except for *mental health* in the general corpus, as shown in Figure 6. Inspecting the top ten ranked collocates for the main target terms (see Appendix F) shows the presence of only two of the 17 pathology-related terms in psychology and the general corpus ("disorder" and "treatment"), and no pathology-related terms among the top ranked collocates for the control. The diversity of terms among the top ranked collocates for *mental health* and *mental illness* indicate that more themes are present in the semantic space.

Salience: Figure 7 illustrates that the relative frequencies rise significantly for both target concepts, *mental health* and *mental illness*, in both corpora. The relative frequency of *perception* increases significantly in the psychology corpus and shows relatively stability in the general corpus.

The significance of the trends was determined by examining standardized beta coefficients and their associated standard errors (see Table 17). As shown in Appendix G, the strongest effect sizes can be observed for the two target terms with breadth (both

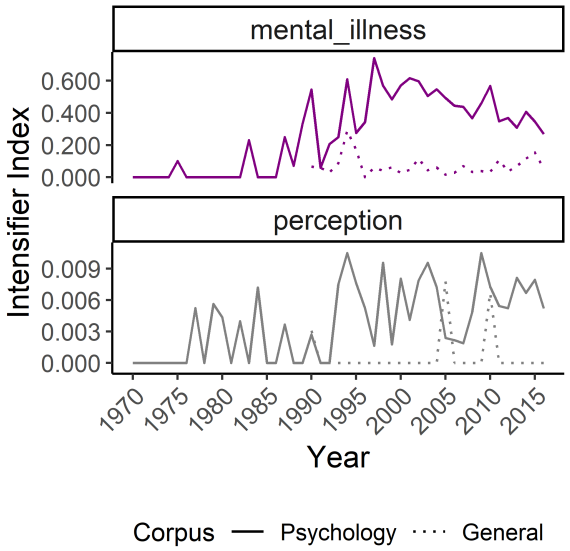


Figure 4: Intensifier index for *mental illness* over the study period (1970-2016).

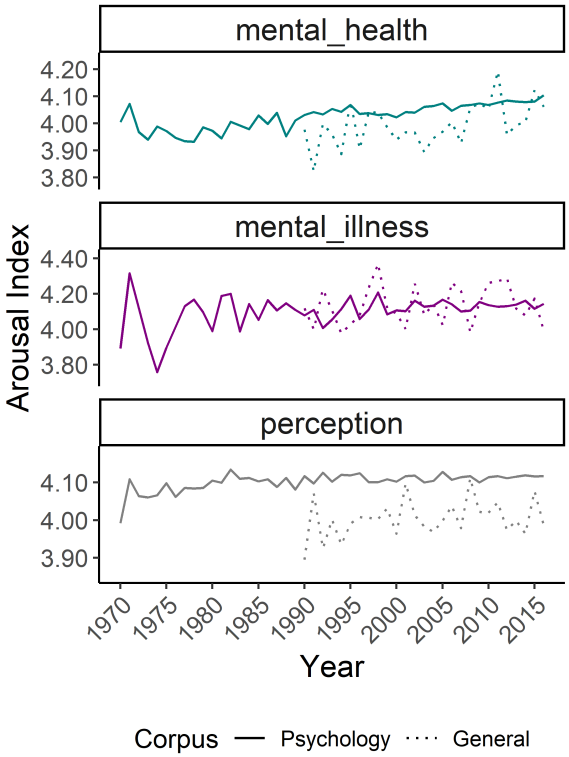


Figure 5: Arousal index over the study period (1970-2016).

corpora), valence (decreasing for psychology and increasing for the general corpus), and for *mental illness* with intensity (both corpora). According to the Adjusted R^2 values in Tables 15 and 16, with a few exceptions year has more explanatory power predicting the semantic indices for the target concepts than for the control concept.

6. Discussion

The present study implemented, for the first time, a new framework for evaluating lexical semantic change. Rather than assessing a single dimension of change or classifying it into a specific taxonomic category, the framework enables the concurrent evaluation of multiple dimensions of semantic change, each corresponding to a well-established dimension of referential or affective meaning. Evaluating semantic change along these dimensions simultaneously allows complex patterns of change to be disentangled and characterized, with possible applications in social science research.

The case study demonstrated a suite of computational methodologies for evaluating the framework's dimensions of change in an examination of *mental health* and *mental illness* motivated by social scientific research questions. Theorists working in sociology, psychology, psychiatry, and related fields have speculated on recent cultural shifts in these concepts, relying on overlapping and sometimes ill-defined notions of medicalization (Bröer and Besseling, 2017; Hofmann, 2016), pathologization (Brinkmann, 2016; Frances, 2013), psychiatrization (Beeker et al., 2021; Paris, 2020), and stigmatization (Sartorius, 2007; Schomerus et al., 2022). Little research has investigated these proposed trends or attempted to characterize them systematically. Our case study documents how a rigorous characterization of these conceptual changes might be conducted. Its findings point to the complexity of these changes, which would remain hidden had they been evaluated on a single dimension.

Regarding sentiment, we found paradoxical trends. Sentiment toward *mental illness* became more positive in the general corpus, supporting suggestions of destigmatization in the culture at large (e.g., Schomerus et al. 2022), while sentiment toward *mental health* and *mental illness* became more negative in the psychology corpus and for *mental health* in the general corpus. In the general corpus, *mental health* came to be used in narrower contexts. Nevertheless, the consistent rising trends for semantic breadth in the psychology corpus support previous claims of expanding meanings or horizontal concept creep (Brinkmann, 2016; Horwitz and Wakefield, 2007, 2012) in academic psychology.

Furthermore, the analysis of intensity yielded clear patterns of change. The target concepts rose

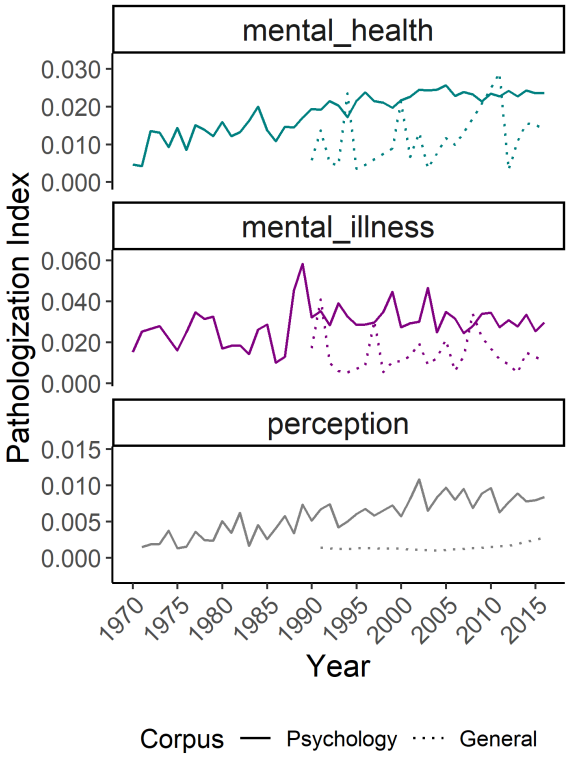


Figure 6: Pathologization index over the study period (1970-2016).

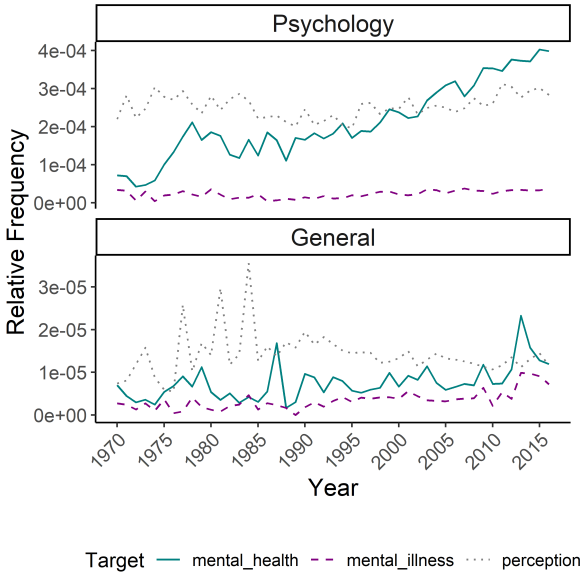


Figure 7: Normalized term frequencies for the general and psychology corpora (1970-2016).

on the arousal index in psychology, indicating that although only the semantic contexts for *mental health* increased in valence, both *mental health* and *mental illness* have become more emotionally animated or agitated. Only *mental health* showed no arousal trend in the general corpus. There was also evidence that *mental illness* has increasingly become more and then less modified by intensifier adjectives, in the psychology corpus, possibly in response to vertical concept creep, where the concept's meaning is stretched to refer to less severe phenomena (Haslam, 2016) which may lead people to intensify the target concept (e.g., where *mental illness* comes to be modified as "serious" or "severe") to distinguish it from more expansive usages. This increase in severity modifiers and arousal may both reflect the same rising concern with and problematization of mental illness and health.

Finally, the tendency for the target concepts to become more associated with pathology-related terms (apart from for *mental health* in the general corpus) supports claims of rising pathologization (Brinkmann, 2016). Notably, *mental illness* was most pathologized. Furthermore, the increase in relative frequency of the target concepts in both corpora is evidence of their rising cultural salience in psychology and the general domain. All indices for the control target, *perception*, showed significant trends in at least one corpus.

In sum, the multi-dimensional analysis suggests that in recent decades, as discourse on mental health and illness has become more prominent (supported by our salience index), concepts of mental health and illness have not so much de-stigmatized (sentiment) but have instead inflated (breadth) and become a growing focus of social concern and problematization (intensity) and increasingly seen through a medical lens (pathologization).

7. Conclusion

The current study presented a new computational framework that can be applied in the social sciences. Our contributions lie in (1) proposing a multidimensional framework to evaluate lexical semantic change in a way that economically integrates forms identified by historical linguists; (2), developing a set of computational methodologies to evaluate change on the newly proposed semantic dimensions; and (3) illustrating the computational framework by examining how *mental health* and *mental illness* have changed their meanings in two corpora, implying that the concepts are increasingly inflated, problematized and pathologized. The investigation illuminates the complexity of semantic and cultural change and provides new tools for studying them.

8. Limitations

Limitations inspire future directions. The procedures employed in the present study are simply a first implementation of the framework. Future research should refine its computational methodology by enhancing or replacing procedures with more robust or sensitive alterations. While the Warriner norms data we used (i) follows a rigorous and reliable rating procedure, (ii) are highly interpretable and (iii) have high face validity, future work might consider alternative methods in addition to closed-vocabulary approaches (Eichstaedt et al., 2021). The current method could be compared against publicly-available BERT-based models fine-tuned for sentiment analysis (Goworek and Dubossarsky, 2024), the VADER (a rule-based sentiment analysis tool; Hutto and Gilbert, 2014), or other sentiment-emotion lexica (Boyd-Graber et al., 2022; Mohammad, 2018). Ideally, the approach will capture the nuanced sentiment contributions of the target word, which averaging the sentiment of contexts fails to capture (Goworek and Dubossarsky, 2024). Robustness checks should be conducted on new methods by comparing its convergent validity against the existing one to evaluate the extent to which the alternative method correlates when applied to the same dataset. In addition, because the target term's semantic broadening is operationalized as the cosine *dis*similarity of the target's sentential contextual usages, it only differentiates between quantitatively (not qualitatively) different meanings. Future work should introduce more fine-grained follow-up analyses by, for example, identifying hypernymy or using state-of-the-art word in context (WiC) models, like XL-LEXEME (Cassotti et al., 2023), which beats GPT-4 on the WiC task and BERT, mBERT, XLM-R on the graded change detection task (Periti and Tahmasebi, 2024). It should also introduce a diachronic analysis to examine if the target's prototypical meaning has been diluted/intensified.

Additionally, while the present study includes a neutral control term, future work should evaluate how to (semi)automatically identify baseline semantic change in the global corpus (a stability axis), to normalize the semantic change of the target concepts against. A control condition where no change of meaning is expected could also be set up (Dubossarsky et al., 2017) using a chronologically shuffled corpus so that the assumed changes become uniform and any change is an artefact (reflects random "noise", not variation in time). To better capture themes, future work should develop a bottom-up, not a top-down dictionary-based, approach by using topic modeling or clustering contextualized word embeddings (Montariol et al., 2021) and evaluating the target's proximity to the centroid of the semantic category cluster. These methods might reveal senses or domains without imposing a dictionary on the semantic space. It will also be crucial to consider LLM approaches for lexical semantic change (Wang and Choi, 2023).

With regard to substantive studies, it will be important to make a general case for the framework by, ideally, finding an existing data set that includes annotated examples of semantic change for evaluation and estimation of the recall/coverage of the methods. In addition, our findings should be extended by applying the framework to a wider assortment of mental health-related concepts such as diagnostic terms (e.g., anxiety, depression, autism, obsessive-compulsive disorder, schizophrenia, attention-deficit hyperactivity disorder). Characterizing how specific diagnoses have altered their meanings in a differentiated, multi-dimensional manner will illuminate historical changes that have only been the focus of theoretical speculation and qualitative research to date (e.g., Brinkmann, 2016; Horwitz and Wakefield, 2007, 2012; Parrott, 2023). Future research can also capitalize on the new framework to explore possible causal relationships between dimensions, such as whether rising salience drives conceptual broadening (Haslam et al., 2021), whether rising breadth of mental illness-related concepts drives improvements in sentiment (a destigmatization process), and whether trade-offs exist (e.g., rising breadth may lead to shifts in intensity). Studies already point to related laws of semantic change, finding that sentiment change is associated with semantic change (Goworek and Dubossarsky, 2024). Future studies should conduct fine-grained analyses on semantic shifts in discourse around mental health to examine how online group dynamics and macro social and cultural shifts (e.g., prevailing stereotypes and stigma towards social groups; see Garg et al., 2018; Charlesworth and Hatzenbuehler, 2024; Durrheim et al., 2023) contribute to observed semantic shifts and possibly the social transmission of mental disorders, shown in adolescent peer networks; Alho et al. (2024). Ideally studies will be conducted with many corpora (e.g., news, social media) with high frequencies of the target terms.

9. Ethics Statement

We do not identify any foreseeable risks or potential for harmful use of our work. Analyses use licensed data that are openly accessible for academic purposes, ensuring transparency and accountability.

Acknowledgements

We thank the three anonymous ACL reviewers for their valuable feedback which substantially improved the paper. Our gratitude also extends to Professor Charles Kemp and Professor Yoshihisa Kashima for their guidance and feedback on earlier versions of the framework at PhD committee meetings and to Lea Frermann, Filip Miletić and Andrey Kutuzov for indirect recommendations which benefited the work, and to Zheng Wei Lim for helping me troubleshoot the airXiv submission. This research is supported by Australian Research Council Discovery Project DP210103984 and by an Australian Government Research Training Program Scholarship.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Janyce Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 19.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Jussi Alho, Mai Gutvilig, Ripsa Niemi, Kaisla Komulainen, Petri Böckerman, Roger T Webb, Marko Elovainio, and Christian Hakulinen. 2024. Transmission of mental disorders in adolescent peer networks. *JAMA psychiatry*.

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 388–401, Toronto, Canada. Association for Computational Linguistics.

Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2023a. Semantic shifts in mental health-related concepts. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 119–128, Singapore. Association for Computational Linguistics.

Naomi Baes, Ekaterina Vylomova, Michael J. Zyphur, and Nick Haslam. 2023b. The semantic inflation of "trauma" in psychology. *Psychology of Language and Communication*, 27(1):23–45.

Timo Beeker, China Mills, Dinesh Bhugra, Sanne te Meerman, Samuel Thoma, Martin Heinze, and Sebastian von Peter. 2021. Psychiatrization of society: A conceptual framework and call for transdisciplinary research. *Frontiers in Psychiatry*, 12:645556.

Andreas Blank. 1999. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. In Andreas Blank and Peter Koch, editors, *Historical semantics and cognition*, pages 61–90. Mouton de Gruter.

Leonard Bloomfield. 1933. *Language*. Compton Printing Works Ltd.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32, Seattle, United States. Association for Computational Linguistics.

Svend Brinkmann. 2016. *Diagnostic Cultures: A Cultural Approach to the Pathologization of Modern Life*. Routledge.

Michel Bréal. 1897. *Essai de sémantique*. Hachette.

Christian Bröer and Broos Besseling. 2017. Sadness or depression: Making sense of low mood and the medicalization of everyday life. *Social Science & Medicine*, 183:28–36.

Lyle Campbell. 1999. *Historical linguistics: An introduction*, 1st mit press ed edition. MIT Press. Available online at `http://tscheer.free.fr/scan/Campbell%2098%20-%20Historical%20Linguistics.%20An%20Introduction.pdf`.

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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and

crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Tessa ES Charlesworth and Mark L Hatzenbuehler. 2024. Mechanisms upholding the persistence of stigma across 100 years of historical text. *Scientific Reports*, 14(1):11069.

Brodie C Dakin, Melanie J McGrath, Joshua J Rhee, and Nick Haslam. 2023. Broadened concepts of harm appear less serious. *Social Psychological and Personality Science*, 14(1):72–83.

Mark Davies. 2008. The corpus of contemporary american english (COCA). https://www.english-corpora.org/coca/.

Mark Davies. 2010. The corpus of historical american english (coha). Available online at https://www.english-corpora.org/coha/.

Simon De Deyne, Daniel J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in characterization of semantic change. *arXiv preprint arXiv:2402.19088*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 4171–4186.

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. A computational exploration of pejorative language in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Paul Dudgeon. 2017. Some improvements in confidence intervals for standardized regression coefficients. *Psychometrika*, 82:928–951.

Kevin Durrheim, Maria Schuld, Martin Mafunda, and Sindisiwe Mazibuko. 2023. Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, 62(1):617–629.

Johannes C Eichstaedt, Margaret L Kern, David B Yaden, H Andrew Schwartz, Salvatore Giorgi, Gregory Park, Courtney A Hagan, Victoria A Tobolsky, Laura K Smith, Anneke Buffone, et al. 2021. Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, 26(4):398.

Lauren Fonteyn and Enrique Manjavacas. 2021. Adjusting scope: A computational approach to case-driven research on semantic change. In *CHR*, pages 280–298.

Allen Frances. 2013. *Saving Normal: An Insider's Revolt Against Out-of-Control Psychiatric Diagnosis, DSM-5, Big Pharma, and the Medicalization of Ordinary Life*. HarperCollins Publishers (Australia) Pty. Ltd., Level 13, 201 Elizabeth Street, Sydney, NSW 2000, Australia.

Frank Furedi. 2016. The cultural underpinning of concept creep. *Psychological Inquiry*, 27(1):34–39.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Dirk Geeraerts. 2010. *Theories of lexical semantics*. Oxford University Press.

Roksana Goworek and Haim Dubossarsky. 2024. Toward sentiment aware semantic change analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 350–357.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 2116. NIH Public Access.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

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

Nick Haslam and Naomi Baes. 2024. What should we call mental ill health? historical shifts in the popularity of generic terms. *PLOS Ment Health*, 1(1).

Nick Haslam, Brodie C Dakin, Fabian Fabiano, Melanie J McGrath, Joshua Rhee, Ekaterina Vylomova, Morgan Weaving, and Melissa A Wheeler.

2020. Harm inflation: Making sense of concept creep. *European Review of Social Psychology*, 31(1):254–286.

Nick Haslam, Ekaterina Vylomova, Michael J. Zyphur, and Yoshihisa Kashima. 2021. The cultural dynamics of concept creep. *American Psychologist*, 76(6):1013–1026.

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

Bjørn Hofmann. 2016. Medicalization and overdiagnosis: Different but alike. *Medicine, Health Care and Philosophy*, 19(2):253–264.

Allan V. Horwitz and Jerome C. Wakefield. 2007. *The loss of sadness: How psychiatry transformed normal sorrow into depressive disorder*. Oxford University Press.

Allan V. Horwitz and Jerome C. Wakefield. 2012. *All we have to fear: Psychiatry's transformation of natural anxieties into mental disorders*. Oxford University Press.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 216–225.

Andrew T. Jebb, Louis Tay, Wei Wang, and Qiming Huang. 2015. Time series analysis for psychological research: Examining and forecasting change. *Frontiers in Psychology*, 6.

Daniel Jurafsky and James H. Martin. 2023. *Vector Semantics and Embeddings*. Draft of February 3, 2024. Draft chapters available online: `https://web.stanford.edu/~jurafsky/slp3/`.

Li Kong, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. 2020. Identifying exaggerated language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7024–7034, Online. Association for Computational Linguistics.

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

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

Lev Lafayette, Greg Sauter, Linh Vu, and Bernard Meade. 2016. Spartan performance and flexibility: An hpc-cloud chimera. *OpenStack Summit, Barcelona*, 27:6.

David E. Levari, Daniel T. Gilbert, Timothy D. Wilson, Baruch Sievers, David M. Amodio, and Thalia Wheatley. 2018. Prevalence-induced concept change in human judgment. *Science*, 360(6396):1465–1467.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yu Luo, Dan Jurafsky, and Beth Levin. 2019. From insanely jealous to insanely delicious: Computational models for the semantic bleaching of english intensifiers. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 1–13.

Christopher D Manning. 2022. Human language understanding & reasoning. *Daedalus*, 151(2):127–138.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Stefano Montanelli and Fabio Periti. 2023. A survey on contextualised semantic shift detection. *arXiv*, arXiv:2304.01666.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652.

Charles Egerton Osgood, William H May, and Murray S Miron. 1975. *Cross-Cultural Universals of Affective Meaning*. University of Illionois Press.

Joel Paris. 2020. *Overdiagnosis in Psychiatry: How Modern Psychiatry Lost Its Way While Creating a Diagnosis for Almost All of Life's Misfortunes*. Oxford University Press.

Scott Parrott. 2023. PTSD in the news: Media framing, stigma, and myths about mental illness. *Electronic News*, 17(3):181–197.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. *arXiv preprint arXiv:2402.12011*.

Ivan Jacob Agaloos Pesigan, Rong Wei Sun, and Shu Fai Cheung. 2023. betadelta and betasandwich: Confidence intervals for standardized regression coefficients in r. *Multivariate Behavioral Research*, 58(6):1183–1186.

Steven Pinker. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Books.

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.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Norman Sartorius. 2007. Stigma and mental health. *The Lancet*, 370(9590):810–811.

Nina Schneidermann, Daniel Hershcovich, and Bolette Pedersen. 2023. Probing for hyperbole in pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 200–211, Toronto, Canada. Association for Computational Linguistics.

Georg Schomerus, Stephanie Schindler, Christian Sander, Eva Baumann, and Matthias C Angermeyer. 2022. Changes in mental illness stigma over 30 years–improvement, persistence, or deterioration? *European Psychiatry*, 65(1):e78.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Cass R. Sunstein. 2018. The power of the normal. *SSRN*. SSRN Scholarly Paper ID 3239204. Social Science Research Network. https://doi.org/10.2139/ssrn.3239204.

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, Berlin.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Yufei Tian, Arvind Krishna Sridhar, and Nanyun Peng. 2021. Hypogen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593.

Jesse S. Y. Tse and Nick Haslam. 2021. Inclusiveness of the concept of mental disorder and differences in help-seeking between asian and white americans. *Frontiers in Psychology*, 12. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.699750.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2022. Comparison and combination of sentence embeddings derived from different supervision signals. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 139–150, Seattle, Washington. Association for Computational Linguistics.

Stephen Ullmann. 1962. *Semantics: An Introduction to the Science of Meaning*. Blackwell.

Ekaterina Vylomova and Nick Haslam. 2021. Semantic changes in harm-related concepts in english. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yue Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*. Language Science Press.

Ekaterina Vylomova, Sean Murphy, and Nick Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, Berlin, Germany. Association for Computational Linguistics.

Ruiyu Wang and Matthew Choi. 2023. Large language models on lexical semantic change detection: An evaluation. *arXiv preprint arXiv:2312.06002*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Melissa A Wheeler, Melanie J McGrath, and Nick Haslam. 2019. Twentieth century morality: The rise and fall of moral concepts from 1900 to 2007. *PLOS ONE*, 14(2):e0212267.

WHO. 2021. Comprehensive mental health action plan 2013–2030.

Yu Xiao, Naomi Baes, Ekaterina Vylomova, and Nick Haslam. 2023. Have the concepts of 'anxiety' and 'depression' been normalized or pathologized? a corpus study of historical semantic change. *PLOS ONE*, 18(6):e0288027.

Arda Yüksel, Berke Uğurlu, and Aykut Koç. 2021. Semantic change detection with gaussian word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3349–3361.

A. Appendix A

To elaborate on what a word being low or high "arousal" or "valence" means, Warriner et al. (2013) defined them in the following way when (valid) participants made direct judgements of the large sample of words on the measured attributes (n = 419: valence; n = 448: arousal; 16-87 years; majority were female (60%), English native language speakers, held a college degree):

- **Valence**: *"You are invited to take part in the study that [...] concerns how people respond to different types of words. You will use a scale to rate how you felt while reading each word. [...] The scale ranges from 1 (happy) to 9 (unhappy). At one extreme of this scale, you are happy, pleased, satisfied, contented, hopeful. When you feel completely happy you should indicate this by choosing rating 1. The other end of the scale is when you feel completely unhappy, annoyed, unsatisfied, melancholic, despaired, or bored. You can indicate feeling completely unhappy by selecting 9. The numbers also allow you to describe intermediate feelings of pleasure, by selecting any of the other feelings. If you feel completely neutral, neither happy nor sad, select the middle of the scale (rating 5)."*

- **Arousal**: *"You are invited to take part in the study that [...] concerns how people respond to different types of words. You will use a scale to rate how you felt while reading each word. [...] The scale ranges from 1 (excited) to 9 (calm). At one extreme of this scale, you are stimulated, excited, frenzied, jittery, wide-awake, or aroused. When you feel completely aroused you should indicate this by choosing rating 1. The other end of the scale is when you feel completely relaxed, calm, sluggish, dull, sleepy, or unaroused. You can indicate feeling completely calm by selecting 9. The numbers also allow you to describe intermediate feelings of calmness/arousal, by selecting any of the other feelings. If you feel completely neutral, not excited nor at all calm, select the middle of the scale (rating 5)."*

B. Appendix B

Total lines where target term appears in the text for both corpora (1970-2016): for the General corpus: mental_health = 3,233; mental_illness = 1,559, perception = 9,440; for the Psychology corpus (1970-2016): mental_health = 26,482; mental_illness = 4,219, perception = 54,694.

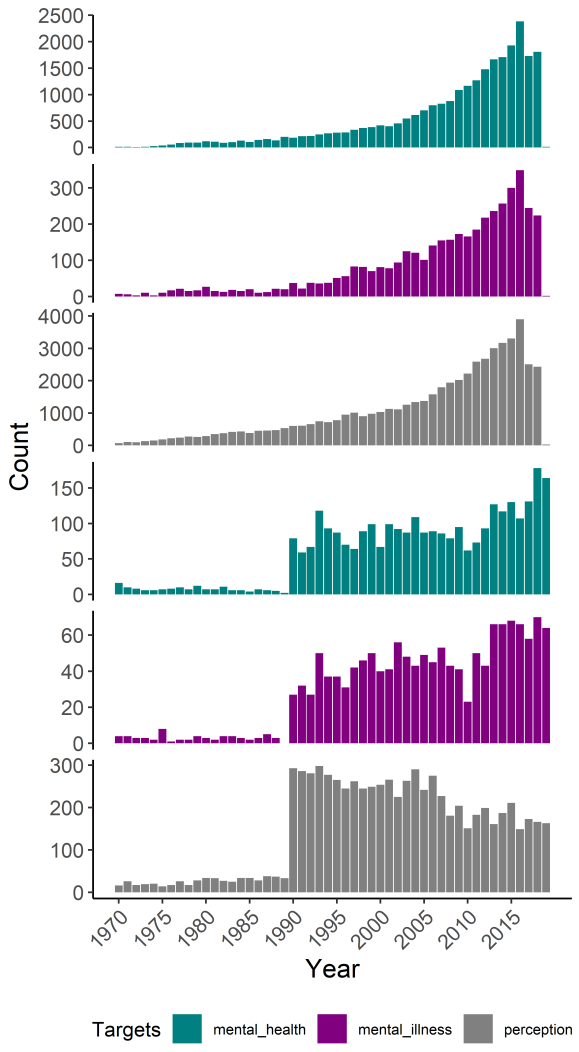


Figure 8: Annual counts of articles where target terms appear in the main text (1970-2016). *Note:* Top three panels = Psychology corpus; bottom three panels = General corpus.

C. Appendix C

Breadth Model Selection

The top three (pre-trained) sentence transformer models were chosen, ranked by their performance in embedding sentences.[11] The best-performing model on the semantic textual similarity benchmark,[12] Multi-Task Deep Neural Network (Liu et al., 2019), was unavailable.[13] See Table 2 for descriptive statistics of models.

- "**all-mpnet-base-v2**"[14] is maintained by the SentenceTransformers community and excels in encoding sentences across 14 diverse tasks from different domains using the MPNet (Masked and Permuted Pre-training for Language Understanding) (Song et al., 2020) architecture.
- "**all-distilroberta-v1**"[15] uses a distilled version of "distilroberta-base" (Sanh et al., 2019), based on BERT architecture, employing knowledge distillation during pre-training and a triple loss (language modeling, distillation and cosine-distance losses) to leverage the inductive biases of LLMs during pre-training.
- "**all-MiniLM-L6-v2**"[16] uses the MiniLM architecture (Wang et al., 2020) employing deep self-attention distillation (using self-attention relation distillation for task-agnostic compression of pre-trained Transformers).
- Additionally, "**bert-base-uncased**"[17] (Devlin et al., 2019) was included for comparison, although its network structure prohibits the direct comparison of sentence embeddings, and BERT maps sentences to a vector space that is unsuitable for use with common similarity measures and performs below average GloVe embeddings on STS tasks (Reimers and Gurevych, 2019).

[11] https://www.sbert.net/docs/pretrained_models.html

[12] https://paperswithcode.com/sota/semantic-textual-similarity-on-sts-benchmark

[13] See https://github.com/namisan/mt-dnn

[14] "all-mpnet-base-v2" from Hugging Face, sentence-transformers: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[15] "all-distilroberta-v1" from Hugging Face, sentence-transformers: https://huggingface.co/sentence-transformers/all-distilroberta-v1

[16] "all-MiniLM-L6-v2": https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[17] "bert-base-uncased": https://huggingface.co/google-bert/bert-base-uncased

Model Info	**all-mpnet-base-v2***	**all-distilroberta-v1***	**all-MiniLM-L6-v2***	**bert-base-uncased**
Accuracy+	69.57	68.73	68.06	NA
Size	420 MB	290 MB	80 MB	80 MB
Case Sensitive	Yes	Yes	Yes	Yes
Vocabulary	30,527	50,264	30,522	30,522
Max Seq Length	384	512	256	512
Pooling	Mean Pooling (Tokens)			CLS pooling
Dimensions	768	768	384	768
Layers	12	6	6	12
Heads	12	12	12	12
Parameters	33M	82.1M	33M	110M
Training Data	>1B training pairs, sent. (3 data sets: wikihow, code_search _net, ms_marco)			NA
Fine-tuning	*Contrastive Learning Objective*: given a sentence from the sentence pair, the model is trained to predict which out of a set of randomly sampled other sentences, is paired with it in the dataset. It computes the cosine similarity from each possible sentence pair and applies the cross-entropy loss by comparing with true pairs.			NA
Base Model	mpnet-base	distilroberta-base	MiniLM-L12-H384-uncased	bert-base-uncased
Pre-training Corpora	BooksCorpus, CC-News, English Wikipedia, OpenWebText, Stories	BooksCorpus, CC-News, English Wikipedia, OpenWebText, Stories	Unknown (Corpora for the original model used for distillation, UniLMv2, is also unknown)	Unknown (Likely a large code and text dataset)
Pre-training Technique	(1) Permuted language modeling; (2) Incorporate auxiliary positional information	(1) Knowledge Distillation, building on the robust training techniques of RoBERTa (dynamic masking, large batch sizes, longer training duration)	(1) Distillation (deep self-attention distillation) likely from UniLMv2	(1) Masked language modeling; (2) Next sentence prediction; (3) Tokenization with WordPiece; (4) Positional embeddings

Table 2: Summary of language models sampled in the present study. *Note:* * = embeddings are normalized. + = Average performance on encoding sentence over 14 tasks over 14 diverse tasks from different domains (14 datasets). SNL = 570k sentence pairs annotated with labels. Multi-Genre NLI = 430k sentence pairs covering spoken and written text. BookCorpus = 11,038 unpublished books scraped from the Internet.

Model Comparison: Test Sample

First, we compared similarity scores for sentence embedding pairs for each sentence transformer model to get a qualitative understanding of the captured dimensions. After feeding seven sample sentences through each sentence transformer model for encoding, similarity arrays of each sentence embedding pair were compared. Tokenization and preprocessing is handled as part of the sentence transformers library.

- 0 = "*She has been seen at a mental_health facility since 1983.*"
- 1 = "*I didn't want to believe I had any mental_health issues and went into denial.*"
- 2 = "*The burden of mental_illness concentrates in 5-10 of the adolescent population.*"
- 3 = "*Their rates of mental_illness are almost twice that of religious adolescents raised in religious households.*"
- 4 = "*Stigma against people with mental_illness is a very complex public health problem.*"
- 5 = "*Stigma associated with mental_illness is one of the major impediments in evolving effective treatment interventions to address the burden associated with these disorders.*"
- 6 = "*Anorexia is a killer it has the highest mortality rate of any mental_illness, including depression .*"

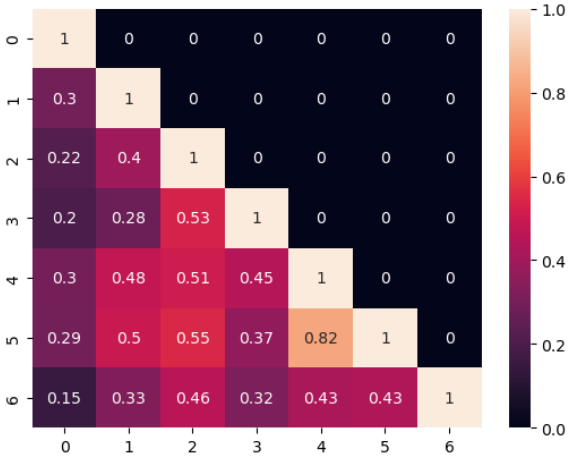


Figure 9: Cosine similarity matrix for sentence embeddings using the "all-mpnet-base-v2" model.

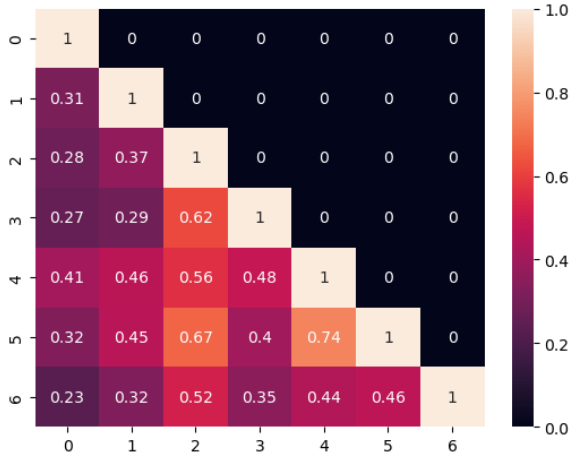


Figure 10: Cosine similarity matrix for sentence embeddings using the "all-distilroberta-v1" model.

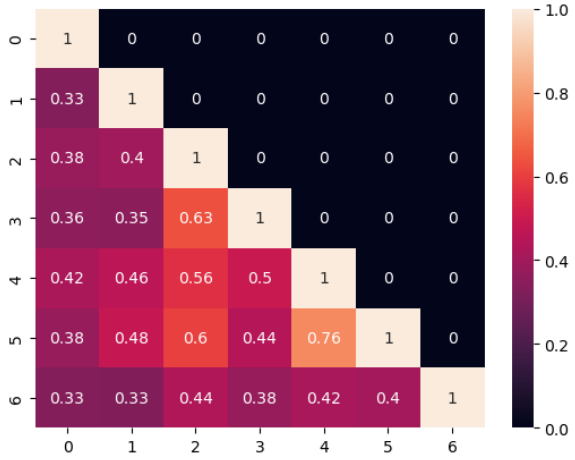


Figure 11: Cosine similarity matrix for sentence embeddings using the "all-MiniLM-L6-v2" model.

Our analysis demonstrated that "all-mpnet-base-v2" (the best model on various encoding tasks, as shown in the first row of Table 2) had the highest similarity for semantically equivalent sentences (see Figure 9). For this task, its superior performance might be attributed to its architecture. MPNet leverages token dependencies through permuted language modeling, which involves scrambling sentential word order and training the model to predict the original order, forcing MPNet to learn the relationships and dependencies between words. It also incorporates auxiliary positional information, allowing the model to perceive entire sentences, enhancing its ability to capture semantic nuances. "all-distilroberta-v1" (Figure 10) and "all-MiniLM-L6-v2" (Figure 11) do not capture this semantic depth, underscoring the strengths of MPNet in semantic understanding and syntactic sensitivity.

Breadth Measure

To analyze semantic differences among sentences containing target concepts, we first extracted texts and then sentences containing target terms from our corpora. The frequency of these sentences over five-year intervals dictated the minimum acceptable number of sentences to sample.

Next, we engaged in a randomized sampling process. In the 1975-1979 interval, there were more than 50 texts in total, apart from for "mental_health", in the general corpus. From these texts, we randomly sampled up to 50 sentences per interval across 10 sets of samples (we sampled all available sentences when there were fewer texts), resulting in up to 500 sentences for every five-year interval, shown in Figure 12.

Following data acquisition, we encoded sentence embeddings using state-of-the-art approaches. Using sentence transformers (except for "bert-base-uncased" which tokenized and passed sentences through PyTorch tensors), we derived embeddings that encapsulated the semantic essence of each sentence. These embeddings were averaged along dimension one in the last hidden state layer, creating a single vector representation for each sentence that achieves a nuanced representation of the sentence's semantic content.

Finally, dissimilarity scores were computed. Leveraging the inverse cosine distance metric, we estimated the similarity between every pair of sentence representations using pairwise distances within the range [-1,1]. To ensure unbiased results, we excluded self-similarity and symmetric elements, focusing solely on the upper half of the matrix (49x25). During analysis, the matrix was flattened to extract a 1D array (a stacked half-matrix) of line-by-line similarity scores. Next, we inverted the similarity scores by subtracting them from 1 to obtain absolute values within the range of [0,1], signifying the dissimilarity between corresponding sentence vectors. The final dissimilarity metric was computed by averaging scores within each of the ten samples per interval (getting the sum of cosine distance scores divided by the total number of sentence pairs), followed by an additional averaging across each five-year period within the 1970-2014 range. Higher scores on the cosine distance metric, ranging from 0 to 1, correspond to greater dissimilarity between sentence vectors.

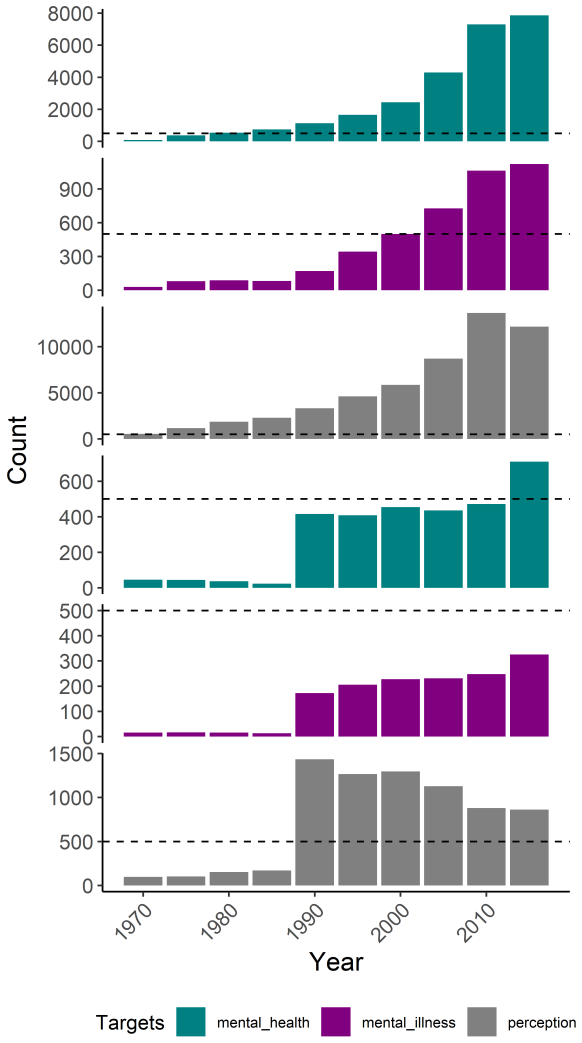


Figure 12: Counts of lines containing target terms grouped by 5-year intervals (horizontal line represents maximum sampling threshold). *Note:* Top three panels = Psychology corpus; bottom three = General corpus.

Model Comparison: Results

After computing breadth scores across five-year intervals, we compared trends using sentence transformer models and bert-base-uncased. As shown in Figure 13, most models showed an upward trend in cosine distance (i.e., inverse similarity), indicating a broader semantic usage of the target concepts. However, "bert-base-uncased" showed lower and flatter similarity scores, possibly due to its pre-training on tasks less directly related to semantic textual similarity. "all-mpnet-base-v2," chosen for the main analysis, performed similarly to the other models but excelled in capturing semantic nuances, as described in Section C.

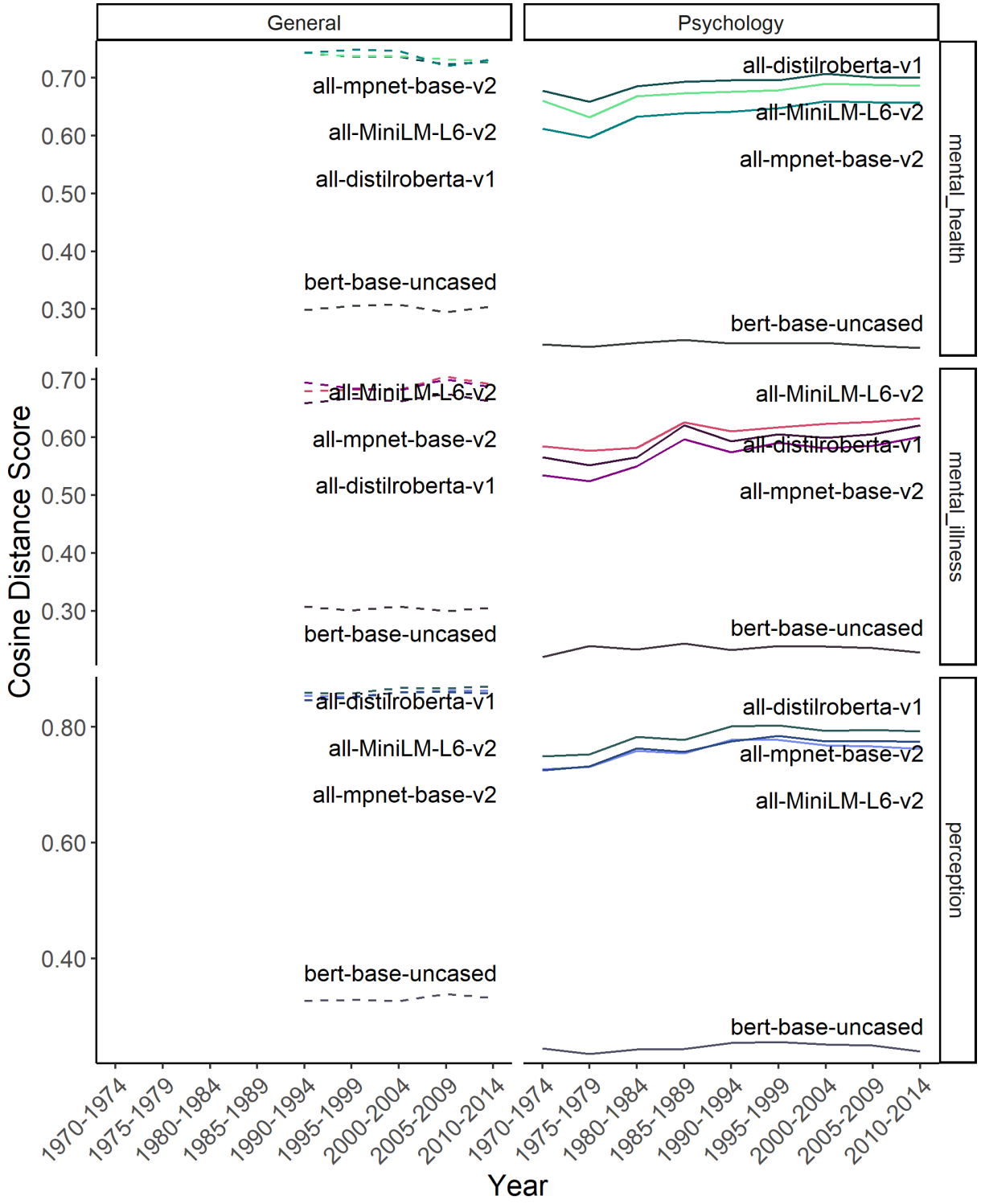


Figure 13: Breadth score over five-year intervals for each model (1970- 2014). *Note:* Model order demonstrates rank of cosine distance score at the final data point (2010-2014) from highest to lowest.

D. Appendix D

To create the general corpus, a rigorous procedure was followed. We first combined two related corpora: the Corpus of Historical American English (CoHA; Davies, 2008) and the Corpus of Contemporary American English (CoCA; Davies, 2008). CoHA contains 400 million words from 1810-2009, drawn from 115,000 texts distributed across everyday publications (fiction, magazines, newspapers, and non-fiction books). CoCA contains 560 million words from 1990-2019 drawn from 500,000 texts (from spoken language, TV shows, academic journals, fiction, magazines, newspapers, and blogs). After merging the two corpora, the combined corpus spanning 1810-2019 was processed following recommendations from Alatrash et al. (2020) to clean it without compromising the qualitative and distributional properties of the data. This process included first excluding the special token "@", which appears in 5% of the CoHA corpus (introduced for legal reasons), malformed tokens that are possible artifacts of the digitization process or the data processing, and clean-up performed using the web interface ("&c?;", "q!", "lp130", "NUL"), and removing escaped HTML characters (" (STAR) ", "<p>", "<>"). Other symbols were excluded after manual inspection of the corpus (e.g., " // ", " | ", " – ", "*", "..", "PHOTO", "(COLOR)", " ILLUSTRATION ", "/"). Blogs were also excluded (89,054 web articles; 98,788 blogs) for not containing associated year data, and 25,418 academic texts were removed. Forty-one lines were removed for missing text data (3 fiction, 11 news, 25 magazines, 2 spoken text) and 32 lines were removed for column misalignment (15 mag, 15 news, 1 fiction, 1 tv). The cleaned corpus was then lowercased and punctuation (commas, periods, question marks), function words, numerals and academic texts were removed. The final combined corpus contained 822,620,111 words from 344,634 texts: 30,496 fiction books, 136,476 magazines, 113,421 newspapers, 2,635 non-fiction books, 43,209 spoken language and 18,397 TV shows. The current study restricted the corpus period from 1970 to 2016 using 501,415,577 tokens from 244,552 articles (23,855 fiction; 88,641 magazines; 73,557 news; 1,498 non-fiction; 40,036 spoken; 16,965 TV).

E. Appendix E

1970	1980	1990	2000	2010
positive	poor	poor	poor	poor
poor	general	maternal	maternal	positive
adolescent	well	positive	well	maternal
female	positive	well	positive	well
improved	preventive	adolescent	general	adolescent
overall	good	general	adolescent	parental
preventive	adolescent	good	parental	general
public	maternal	bad	bad	bad
recent	own	optimal	good	good
robust	individual	own	overall	overall

Table 3: Top 10 adjective modifiers of *mental health* in the psychology corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
past	chronic	severe	severe	severe
chronic	major	serious	serious	serious
excess	severe	chronic	major	chronic
feminine	familial	major	chronic	parental
formal	malingered	common	parental	major
more	acute	maternal	other	common
obvious	aggressive	parental	common	other
other	disabling	comorbid	maternal	co
partum	few	persistent	comorbid	maternal
severe	less	other	persistent	comorbid

Table 4: Top 10 adjective modifiers of *mental illness* in the psychology corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
visual	visual	visual	visual	visual
interpersonal	social	social	positive	social
auditory	maternal	positive	negative	negative
differential	positive	negative	social	positive
social	parental	subjective	subjective	conscious
subliminal	negative	parental	conscious	subjective
pictorial	interpersonal	maternal	parental	high
binocular	human	auditory	high	auditory
favorable	subjective	accurate	categorical	low
high	auditory	interpersonal	low	parental

Table 5: Top 10 adjective modifiers of *perception* in the psychology corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
collective	everincreasing	good	good	poor
diminished	vibrant	own	well	good
necessary	NA	positive	own	well
normal	NA	rural	collective	abysmal
NA	NA	sound	optimal	additional
NA	NA	subsequent	poor	comprehensive
NA	NA	bad	postpartum	lessthanoptimal
NA	NA	dubious	collegestudent	new
NA	NA	geriatric	confident	own
NA	NA	maternal	fragile	pediatric

Table 6: Top 10 adjective modifiers of *mental health* in the general corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
serious	acute	severe	severe	serious
certain	hereditary	serious	serious	severe
incipient	severe	major	major	other
socalled	socalled	chronic	other	acute
NA	underlying	untreated	most	chronic
NA	NA	common	adolescent	common
NA	NA	other	bipolar	diagnosable
NA	NA	classic	common	major
NA	NA	more	many	deep
NA	NA	severe	new	difficult

Table 7: Top 10 adjective modifiers of *mental illness* in the general corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
extrasensory	public	public	public	public
visual	different	widespread	common	common
own	common	visual	popular	popular
aesthetic	general	common	wrong	general
direct	innate	popular	human	own
single	own	general	general	sensory
keen	popular	own	own	veridical
new	psychic	new	extrasensory	extrasensory
practical	clear	extrasensory	acute	negative
present	human	human	visual	parental

Table 8: Top 10 adjective modifiers of *perception* in the general corpus (terms are ranked by their relative count for the respective decade)

F. Appendix F

1970	1980	1990	2000	2010
community	service	service	service	service
center	community	child	problem	problem
service	professional	professional	child	child
program	center	use	use	study
professional	problem	problem	care	use
child	use	care	study	care
school	study	study	treatment	treatment
problem	social	treatment	professional	need
group	child	need	need	outcome
worker	program	community	health	physical

Table 9: Top 10 Warriner-matched collocates of *mental health* in the psychology corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
attitude	attitude	severe	severe	people
scale	study	person	people	severe
patient	patient	patient	use	study
study	high	treatment	patient	stigma
group	person	substance	study	use
psychiatric	problem	study	person	individual
opinion	scale	use	disorder	treatment
factor	child	people	treatment	disorder
find	use	disorder	individual	family
student	major	family	substance	experience

Table 10: Top 10 Warriner-matched collocates of *mental illness* in the psychology corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
self	study	study	study	study
study	child	self	self	self
result	self	child	child	social
test	result	social	examine	relationship
visual	subject	result	relationship	use
child	difference	examine	use	examine
subject	social	use	social	influence
group	relationship	relationship	result	effect
difference	effect	difference	relate	result
use	group	behavior	influence	child

Table 11: Top 10 Warriner-matched collocates of *perception* in the psychology corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
department	center	have	say	have
state	institute	national	have	say
center	service	institute	national	issue
health	fund	care	institute	care
city	have	service	child	problem
director	national	professional	care	system
institute	allow	abuse	community	service
national	commissioner	state	need	health
new	department	center	service	physical
program	oak	department	problem	professional

Table 12: Top 10 Warriner-matched collocates of *mental health* in the general corpus (terms are ranked by their relative count for the respective decade)

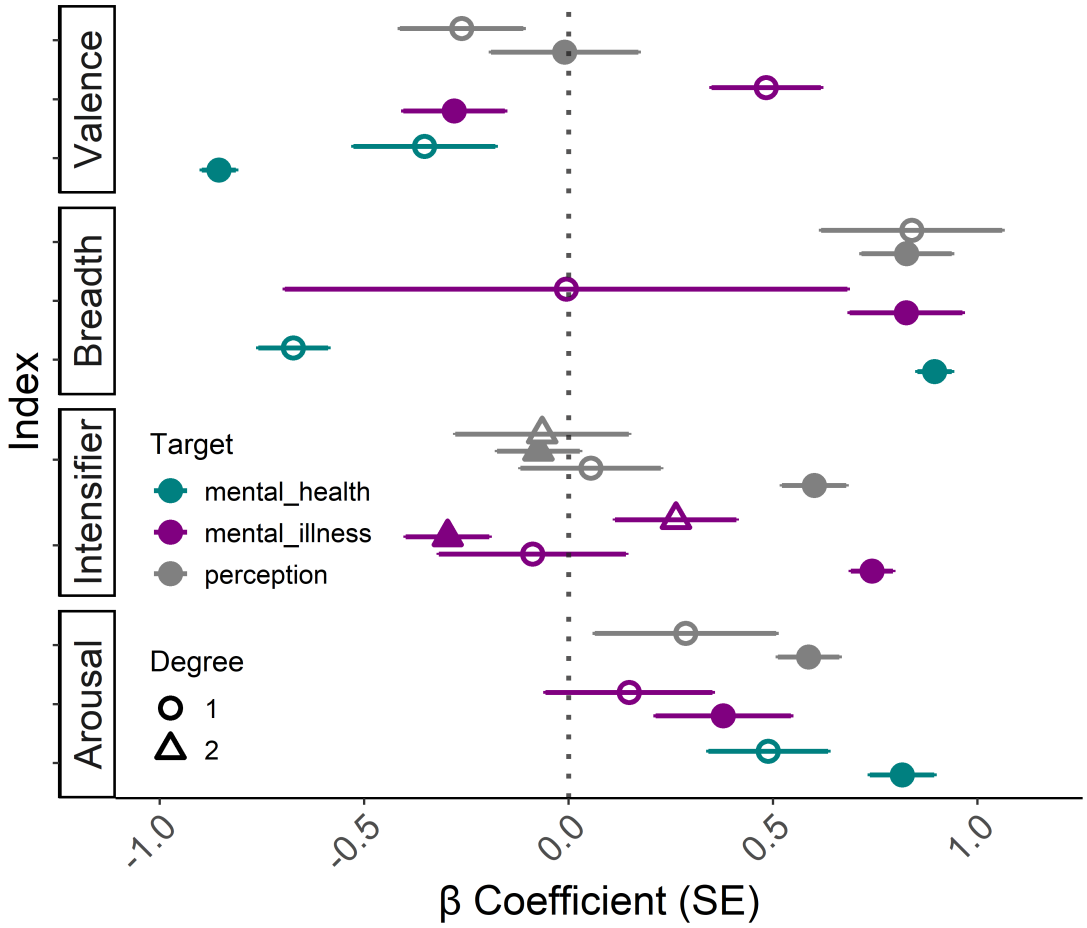
1970	1980	1990	2000	2010
drug	suffer	have	have	have
history	alcoholism	people	people	people
treat	have	severe	family	family
acute	time	depression	do	say
appoint	acute	family	suffer	alliance
bill	argue	say	disorder	history
can	ask	can	say	national
cancer	basement	drug	severe	severe
cause	bout	history	child	suffer
center	cite	know	drug	member

Table 13: Top 10 Warriner-matched collocates of *mental illness* in the general corpus (terms are ranked by their relative count for the respective decade)

1970	1980	1990	2000	2010
people	public	public	change	change
public	change	change	public	public
alter	reality	can	people	people
change	base	people	can	can
president	black	other	reality	time
study	member	reality	other	shift
associate	new	world	go	affect
can	people	may	depth	alter
cause	popular	go	know	challenge
child	side	thing	will	pain

Table 14: Top 10 Warriner-matched collocates of *perception* in the general corpus (terms are ranked by their relative count for the respective decade)

G. Appendix G



Year effect sizes for indices operationalizing major dimensions of lexical semantic change in the psychology corpus (filled circles) and general corpus (empty circles). *Note:* First degree = Linear; Second degree = Quadratic. Vertical dotted line = Standardized beta coefficient of 0; Standard errors (SE) that overlap line indicate that the null hypothesis can be rejected at the 5% significance level.

Index (Concept)	**Corpus**	**Model**	***B***	**SE**	***t***	***p***	***F* (DF); Adj. R^2**
Intensifier (*Mental Illness*)	Psychology	Linear	0.74	0.09	8.18	<.001	38.76 (1, 44); 0.62*
		Quadratic	-0.33	0.10	-3.26	0.002	
	General	Linear	-0.09	0.20	-0.45	.657	0.99 (2, 24); -0.0005
		Quadratic	0.30	0.22	1.34	0.194	
Intensifier (*Perception*)	Psychology	Linear	0.60	0.12	5.00	<.001	12.71 (2,44); 0.34*
		Quadratic	-0.08	0.14	-0.62	0.541	
	General	Linear	0.05	0.20	0.27	0.793	0.08 (2, 24); -0.08
		Quadratic	-0.07	0.23	-0.32	0.752	

Table 15: Regression Coefficients (Scaled) and Fit Statistics Predicting Intensifier Indices as a Function of Year. *Note*: * = *p*-value for the overall model = <.001. Regression coefficients are unstandardized. For mental_illness in psychology, residuals were autocorrelated, and outcome variable was re-fit with Generalized Least Squares approach, yielding: B = 0.74; SE = 0.09; $p < .001$; RSE(DF) = 0.62(47,44); BIC = 108.52.

Index	Concept	Corpus	*B*	SE	*p*	*F* (DF)	Adj. R^2
Valence	*Mental Health*	Psychology	-0.003	3×10^{-4}	<.001	122.65 (1,45)	0.73
		General	-0.005	0.003	.071	3.55 (1,25)	0.09
	Mental Illness	Psychology	-0.002	9×10^{-4}	.057	3.82 (1,45)	0.058
		General	0.01	0.005	.011	7.62 (1,25)	0.20
	Perception	Psychology	-1×10^{-5}	2×10^{-4}	.949	0.004 (1,45)	-0.02
		General	-0.002	0.002	.188	1.84 (1,25)	0.03
Breadth	*Mental Health*	Psychology	0.001	3×10^{-4}	0.001	28.19 (1,7)	0.77
		General	-0.001	7×10^{-4}	.213	2.49 (1,3)	0.27
	Mental Illness	Psychology	0.002	4×10^{-4}	.006	14.99 (1,7)	0.64
		General	-6×10^{-6}	6×10^{-4}	.992	1×10^{-4} (1,3)	-0.33
	Perception	Psychology	0.001	3×10^{-4}	0.006	15.12 (1,7)	0.64
		General	7×10^{-4}	3×10^{-4}	.076	7.13 (1,3)	0.61
Arousal	*Mental Health*	Psychology	0.003	3×10^{-4}	<.001	89.38 (1,45)	0.66
		General	0.005	0.002	<.001	7.83 (1,25)	0.21
	Mental Illness	Psychology	0.003	9×10^{-4}	<.001	7.51 (1,45)	0.12
		General	0.002	0.003	.462	0.56 (1,25)	-0.02
	Perception	Psychology	0.001	2×10^{-4}	<.001	23.65 (1,45)	0.33
		General	0.002	0.001	.148	2.22 (1,25)	0.05
Path.	*Mental Health*	Psychology	4×10^{-4}	3×10^{-5}	<.001	163.34 (1,45)	0.78
		General	3×10^{-4}	2×10^{-4}	.130	2.48 (1,21)	0.06
	Mental Illness	Psychology	2×10^{-4}	1×10^{-4}	.049	4.12 (1,43)	0.07
		General	-1×10^{-4}	2×10^{-4}	.552	0.36 (1,23)	-0.03
	Perception	Psychology	2×10^{-3}	4×10^{-2}	<.001	118.42 (1,44)	0.72
		General	5×10^{-5}	2×10^{-5}	.051	5.95 (1,6)	0.41
Salience	*Mental Health*	Psychology	7×10^{-6}	4×10^{-7}	<.001	292.52 (1,45)	0.86
		General	2×10^{-7}	4×10^{-8}	<.001	18.17 (1,45)	0.27
	Mental Illness	Psychology	3×10^{-7}	9×10^{-8}	<.001	13.21 (1,45)	0.21
		General	1×10^{-7}	2×10^{-8}	<.001	42.21 (1,45)	0.47
	Perception	Psychology	5×10^{-7}	3×10^{-7}	.160	2.04 (1,45)	0.02
		General	-3×10^{-8}	6×10^{-8}	.568	0.33 (1,45)	-0.01

Table 16: Unstandardized Regression Coefficients and Fit Statistics Predicting Indices as a Function of Year. *Note*: The midrule separates the main dimensions (above) and the exploratory dimensions (below). Path. = Pathologization. Generalized Least Squares approach also used for models with autocorrelated residuals.
• Arousal: *mental_health* (P): B = 0.003; SE = 3×10^{-4}; $p < .001$; RSE(DF) = 0.03(47,45); BIC = -172.07
• Salience: *mental_health* (P): B = 7×10^{-6}; SE = 4×10^{-7}; $p < .001$; RSE(DF) = 4×10^{-5}(47,45); BIC = -767.87; *mental_illness* (P): B = 3×10^{-7}; SE = 9×10^{-7}; $p < .001$; RSE(DF) = 9×10^{-6}(47,45); BIC = -895.27; *perception* (P): B = 5×10^{-7}; SE = 3×10^{-7}; $p = .160$; RSE(DF) = 3×10^{-5} (47,45); BIC = -785.60; *mental_illness* (G): B = 1×10^{-7}; SE = 2×10^{-8}; $p < .001$; RSE(DF) = 2×10^{-6}(47,45); BIC = -1048.85

Index	Concept	Corpus	β	SE	95% CI
Valence	*Mental Health*	Psychology	-0.86*	0.04	(-0.94, -0.77)
		General	-0.35	0.17	(-0.71, 0.004)
	Mental Illness	Psychology	-0.28*	0.12	(-0.53, -0.03)
		General	0.48*	0.13	(0.21, 0.76)
	Perception	Psychology	-0.01	0.18	(-0.37, 0.35)
		General	-0.26	0.15	(-0.57, 0.05)
Breadth	*Mental Health*	Psychology	0.90*	0.04	(0.80, 0.99)
		General	-0.67*	0.09	(-0.95, -0.40)
	Mental Illness	Psychology	0.83*	0.14	(0.50, 1.15)
		General	-0.01	0.69	(-2.19, 2.18)
	Perception	Psychology	0.83*	0.11	(0.57, 1.09)
		General	0.84*	0.22	(0.13, 1.54)
Intensifier	*Mental illness*	Psychology(1)	0.74*	0.05	(0.64, 0.85)
		Psychology(2)	-0.30*	0.10	(-0.50, -0.09)
		General(1)	-0.09	0.23	(-0.56, 0.38)
		General(2)	0.26	0.15	(-0.05, 0.57)
	Perception	Psychology(1)	0.60*	0.08	(0.44, 0.76)
		Psychology(2)	-0.07	0.10	(-0.28, 0.13)
		General(1)	0.05	0.17	(-0.30, 0.41)
		General(2)	-0.06	0.21	(-0.50, 0.37)
Arousal	*Mental Health*	Psychology	0.82*	0.08	(0.66, 0.97)
		General	0.49	0.15	(0.19, 0.79)
	Mental Illness	Psychology	0.38*	0.17	(0.05, 0.71)
		General	0.15	0.20	(-0.27, 0.57)
	Perception	Psychology	0.59*	0.08	(0.44, 0.74)
		General	0.29	0.22	(-0.17, 0.74)
Pathologization	*Mental Health*	Psychology	0.30*	0.12	(0.06, 0.53)
		General	-0.12	0.23	(-0.61, 0.36)
	Mental Illness	Psychology	0.89*	0.02	(0.85, 0.92)
		General	0.32	0.20	(-0.09, 0.74)
	Perception	Psychology	0.85*	0.30	(0.79, 0.92)
		General	0.71	0.30	(-0.03, 1.45)
Salience	*Mental Health*	Psychology	0.93*	0.02	(0.89, 0.97)
		General	0.54*	0.10	(0.34, 0.73)
	Mental Illness	Psychology	0.48*	0.13	(0.21, 0.74)
		General	0.70*	0.07	(0.56, 0.83)
	Perception	Psychology	0.21	0.15	(-0.10, 0.52)
		General	-0.09	0.15	(-0.38, 0.21)

Table 17: Standardized Regression Coefficients (β) predicting Semantic Change Indices by Year. *Note*: Midrule separates main dimensions of semantic change (above). * = p: < .05. (1) = First degree. (2) = Second degree.