Chapter 2

Semantic changes in harm-related concepts in English

Ekaterina Vylomova & Nick Haslam
The University of Melbourne

The chapter investigates semantic changes in core concepts of psychology, specifically focusing on those related to harm. Haslam (2016) hypothesized that many psychological concepts associated with harm (i.e., forms of psychological disturbance, threat, and maltreatment) have undergone semantic broadening in the past half-century in association with cultural shifts and social change. The implications of this “concept creep” hypothesis have been previously explored by prominent social, political, and legal thinkers (Levari et al. 2018, Lukianoff & Haidt 2019, Pinker 2018, Sunstein 2018), but its linguistic dimension has received little empirical attention.

Here we apply computational models in order to address the concept creep hypothesis. We start with a description of a typology of semantic shifts and provide a summary of computational methods for automatic detection of the most common changes (broadening, narrowing, hyperbole, and litotes) and utilise those to evaluate core harm-related concepts such as ‘trauma’, ‘harassment’, and ‘bullying’ on a new corpus of psychology literature extending from 1970 to 2017. Our results confirm the initial hypothesis and are in line with earlier studies: most concepts became broader and milder over the last few decades. We then continue with a more detailed study in order to understand how exactly the concepts changed, and to do so employ and evaluate different types of semantic representations.

Finally, we additionally train the models on a general domain corpus in order to investigate whether the broadening of harm-related concepts also applies to society at large, rather than only to the academic discourse of psychology. Haslam’s influential account of concept creep (Haslam 2016) proposes that broadened concepts of harm disseminate from academic language into wider public use. This final analysis enables a direct test of that conjecture, including comparative analysis of the extent and timing of historical semantic changes across the two corpora.

Ekaterina Vylomova & Nick Haslam. 2021. Semantic changes in harm-related concepts in English. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 93–121. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040304

1 Introduction

Recent years witnessed significant progress in many downstream tasks in natural language processing (NLP) such as machine translation, part-of-speech tagging, language modelling, and many others.[1] Unlike earlier machine learning models that were often provided with a set of pre-designed features or rules, most recent models inherently "learn" them from raw data in the form of dense vectors (embeddings). Training strategies used in the models to learn the embeddings often rely on the distributional semantics hypothesis that states that a word's meaning can be expressed as a distribution over a set of its contexts (Firth 1957, Harris 1954, Weaver 1955). A significant amount of research works explored what aspects of language are captured in these representations. Although the distributional semantics approach presents certain limitations (Bender & Koller 2020), it still allows to extract a surprising amount of information about semantic, morphological, and syntactic properties of language (Mikolov, Yih, et al. 2013, Vylomova et al. 2016, Gladkova et al. 2016, Belinkov & Glass 2019, Rogers et al. 2020). In addition, representations obtained using this approach capture associations between words and can potentially simulate surveys on free word associations (Agirre et al. 2009, Antoniak & Mimno 2018). These successes induced a novel direction of interdisciplinary studies – corpus-centered research – where embeddings are used as a direct evidence about the language and culture of the authors of a training corpus (Antoniak & Mimno 2018). For instance, Hamilton et al. (2016a,b) presented one of the earliest diachronic language models and metrics to evaluate semantic shifts as well as computational approaches to lexical semantic change detection. Over the last few years, the area has significantly increased and witnessed substantial progress and development (Schlechtweg et al. 2020).

In this chapter, we apply diachronic language modelling to computationally attest semantic shifts in core concepts of social psychology. In particular, we focus on diachronic change in the meaning of harm-related concepts and test a "concept creep" hypothesis proposed in Haslam (2016). The hypothesis states that during the past half-century many concepts associated with harm have broadened their meanings in Western societies. We quantitatively evaluate changes in the five negative concepts: 'addiction', 'bullying', 'harassment', 'prejudice', and 'trauma'. We attest them on a newly introduced corpus of psychology journal abstracts and a general domain corpus comprising CoCA and CoHA. In order to test the hypothesis, we first conduct frequency-based analysis and then study the changes in a greater detail by evaluating vector representations learned by epoch-specific models trained on each corpus.

[1] See https://nlpprogress.com/ for most recent state-of-the-art models in each task.

2 The notion of concept creep

Haslam (2016) introduced the idea of "concept creep" to describe a general pattern of semantic inflation in several fundamental psychological concepts. The paper presented a series of case studies in which psychological researchers and theorists expanded the sense of harm-related concepts by loosening definitions to include milder instances ("vertical creep") or by extending definitions to encompass qualitatively new phenomena ("horizontal creep").

The two forms of creep can be understood from the perspective of Bloomfield's typology of lexical semantic change (Bloomfield 1933). Out of seven types identified in the book, some of them are particularly relevant to the current creep study. First, changes may happen along the semantic narrowing (the Old English *mete* 'food' > *meat* 'edible flesh') – widening (the Middle English *briddle* 'young birdling' > *bird* 'birds of all ages') axis. Alternatively, a word's meaning may extend by means of analogy (the Old English *bītan* 'to bite' > the Middle English *bitter* 'acrid').

Indeed, modern studies of word semantics change are based on a long tradition. Yet in the end of the 19th century Bréal (1897) analyzed different types of word meaning change in a diachronical perspective for multiple languages. Particularly, the four major types of concept creep discussed in the current chapter (two vertical and two horizontal ones) were reflected in some form within the taxonomy proposed by Bréal. The horizontal concept broadening is similar to what he referred to as "élargissement de sens" (sense enlargement). One of the examples mentions Latin *pecunia* the meaning of which has gradually broadened from 'richness in possession of livestock' to a general sense of 'wealth'. The vertical broadening deems falling into the "épaississement de sens" category ("sense thickening"). We can notice that in the latter case Bréal mostly speaks about facts of meaning change accompanied by either morphological or non-morphological modification of a word in hand. Thus, a word was not required to keep its exact form, in contrast to the approach we follow in the current study. The phenomenon of concept narrowing was not directly outlined in the Bréal's taxonomy. However, both its horizontal and vertical types seem to be covered by different kinds of metaphor.

Horizontal creep comprises both of these types: widening of 'abuse' to include passive neglect and metaphoric extension of (physical) 'bullying' to include 'cyber-bullying'. Another type of shift might occur along the litotes–hyperbole axis. Litotes represents the change from a weaker to a stronger meaning (the Proto-West Germanic **kwalljan* 'to make suffer' > the Old English *cwellan* 'to kill'), whereas hyperbole is the shift in the opposite direction (the Vulgar Latin

**extonare* 'to strike with thunder' > *astonish* 'to surprise greatly'). This type of change seems to be more pertinent to vertical creep: as we will further show, 'trauma' has transformed to refer to relatively mild adversities (Haslam & McGrath 2020). Horizontal and vertical creep are not mutually exclusive – a concept may change in both ways simultaneously. For example, the concept of 'mental disorder' has progressively broadened in recent decades by relaxing the diagnostic criteria of some conditions (vertical creep; Fabiano & Haslam 2020) and by expanding the range of problems conceptualized as falling within the psychiatric domain (horizontal creep).

Haslam (2016) and Haslam et al. (2020) documented how similar semantic inflation had occurred for the following putatively creeping concepts which we will further examine in the current chapter:

Addiction: This concept originally referred to physiological dependency on an ingested substance, but is increasingly used to identify psychological compulsions to engage in non-ingestive behaviors such as gambling or shopping.

Bullying: This concept, introduced to psychology in the 1970s, initially described peer aggression between children that was repeated, intentional, and perpetrated in the context of a power imbalance. More recent definitions extend bullying to adult workplace settings and relax the repetition, intentionality, and power imbalance criteria.

Harassment: Early uses of this concept emphasized inappropriate sexual approaches but more recently harassment is also used within psychology to refer to nonsexual forms of unwanted attention.

Prejudice: The original psychological definitions of prejudice restricted it to overt animosity towards ethnic or racial outgroups. More recent theory and research extend it to many non-racial groups, allow for covert or nonconscious prejudice, and indicate that it may be manifest as anxiety or condescension rather than hostility. Recent studies showed that it expanded to include subtle micro-aggressions (Lilienfeld 2017).

Trauma: Four decades ago only personally encountered life-threatening events that are outside the realm of normal experience were recognized as traumatic by psychologists and psychiatrists. More recent definitions include vicarious or indirect experiences of stressful events, including those that are relatively prevalent.

Haslam (2016) proposed that these diverse concepts shared a focus on harm (i.e., the experience or infliction of actual or potential suffering). It was further speculated that the correlated broadening of the creeping concepts reflected a rising sensitivity to harm within Western cultures.

3 Related work

We will provide related research for three aspects of our study: the central hypothesis of "concept creep", computational approaches to semantic change detection, and factors that might influence semantic change.

3.1 "Concept creep"

Existing work on concept creep with a few notable exceptions is primarily theoretical and the idea has been taken up by influential writers. Lukianoff & Haidt (2019) have employed it to understand political conflict on college campuses. Pinker (2018) has argued that concept creep leads people to under-estimate social progress because their definitions of hardship expand to include increasingly minor problems. This phenomenon has been demonstrated by Levari et al. (2018), who showed that concept definitions broaden as concept instances become scarcer. McGrath et al. (2019) have explored the attributes of people who hold relatively broad harm-related concepts, finding that they tend to be politically liberal and empathetic, and their personal morality is tied to harm and care for others. Wheeler et al. (2019) studied the Google Books English language corpus and showed that words representing harm-based morality has become more culturally salient (i.e., relatively frequent) in the past four decades, consistent with the theory of concept creep. Most recently, Vylomova et al. (2019) trained a count-based model from Sagi et al. (2009) and a prediction-based one introduced in Hamilton et al. (2016b) on a massive corpus of abstracts of academic psychology journals to evaluate semantic breadth changes in some of the creeping concepts described in Haslam (2016).

3.2 Computational approaches to semantic change detection

Although diachronic studies of language have a long history in linguistics, computational approaches were introduced only recently. Jurgens & Stevens (2009), one of the first, proposed an algorithm for tracking temporal semantic changes by learning a sequence of distributional models over time. The work was followed by an LSA-based model from Sagi et al. (2009). Kim et al. (2014) and Hamilton et al. (2016b) then proposed the first prediction-based neural language models. The training strategies of the models differed, though: Kim et al. (2014) incrementally trained models on each subsequent epoch, while Hamilton et al. (2016b) trained several epoch-specific models independently and then aligned them using Procrustes. Kulkarni et al. (2015) also followed the same direction but only aligned the nearest neighbors rather than the whole space. Both Kulkarni et al. (2015) and Hamilton et al. (2016b) further demonstrated that such prediction-based models (word2vec, in particular) outperform count-based ones on the semantic shifts detection tasks. Further, Dubossarsky et al. (2019) demonstrated that alignment-based diachronic models often introduce additional noise to the representations and proposed a temporal referencing approach that does not require vector space alignment.

3.3 Factors that influence semantic changes

Hamilton et al.'s work in 2016 was influential because they also attempted to state laws of semantic change that would explain the variability in word change rates and identify factors that influence said rates. On the other hand, this research direction was not entirely novel for the scientific community outside of NLP: historical linguistics presents a vast line of work on this topic. For instance, Stern (1931) and Lehrer (1985) suggested that words with close meanings that are strongly associated with one another undergo similar changes ("the law of parallel change"). Contrary to that, Sturtevant (1917) stated "the law of differentiation", i.e. that words with similar meanings (synonyms) tend to diverge over time. Xu & Kemp (2015) evaluated the two laws and provided more evidence for support of "the law of parallel change". Geeraerts et al. (1999) suggested that prototypicality also plays a role: more salient, prototypical meanings will be less likely to change. "The law of prototypicality" was then examined in Dubossarsky et al. (2015), the work demonstrating that the closer a word is to the centroid of the corresponding semantic category cluster, the less likely its meaning changes. Another linguistic hypothesis states that "words become semantically extended by being used in diverse contexts" (Winter et al. 2014) and meaning evolves in a directional fashion: words that have more word associations and senses are more likely to acquire new meanings. Finally, Hamilton et al. (2016b) proposed a hypothesis stating that frequency and polysemy explain most variance in the rates of lexical semantic change. Their study resulted in a more comprehensive understanding of the earlier observations, and resulted in the following two laws of semantic change: (1) "The law of conformity": frequently used words change at slower rates; and (2) "The law of innovation": polysemous words change at faster rates. Later, Dubossarsky et al. (2017) re-considered the laws of semantic change and showed that (1) “the law of innovation” is to a large extent an artefact of frequency; (2) “the law of conformity” is also an artefact of word representation models; and (3) the impact of prototypicality proposed in Dubossarsky’s earlier work is smaller.

4 Corpora: Psychology and general domain

In the current study we compare dynamics of concept breadth in two corpora: a corpus of psychology abstracts (domain-specific) and a compilation of the corpus of historical English (CoHA; Davies 2012) and the corpus of contemporary American English (CoCA; Davies 2008) texts (general domain).

4.1 Psychology corpus

The corpus comprises abstracts from journals in the field of psychology covering the period of 1930–2019 that were collected from the E-Research and the PubMed databases. In total, there are 871,340 abstracts from 875 journals resulting in 133,082,240 tokens. We only focus on abstracts since they distill the core ideas of the paper and provide a compact summary of the main contributions and findings.[2] Figure 2.1 presents the number of abstracts for each year. Due to the relatively small amount of abstracts during the first half of the 20th century, for the purpose of our experiments we only consider time periods after 1970. We also exclude two final years (2018, 2019) due to the lack of data from one of the databases.

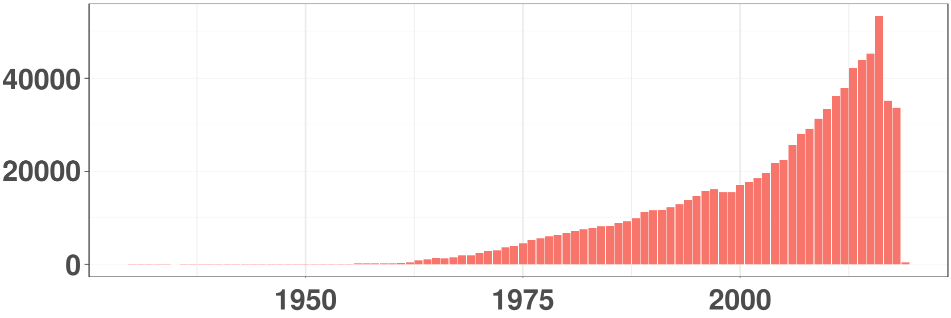


Figure 2.1: Statistics on the number of abstracts per year

[2] Restrictions related to copyright also limited our focus to abstracts.

4.2 CoCA and CoHA

The corpus of historical English (CoHA) starts in the 1810s and ends in the early 2000s, comprising 400 million words from 115,000 texts evenly sampled for each decade from fiction, magazines, newspapers, and non-fiction books.

The corpus of contemporary American English (CoCA) covers the period from 1990 till 2019 and contains about 1 billion words from 500,000 texts evenly sampled from spoken, TV shows, academic journals, fiction, magazines, newspapers, and blogs.

For the purpose of the study, we combined the two corpora leaving only the period between 1970 and 2017. We excluded blogs because of the lack of timestamps and additionally removed texts extracted from academic journals to ensure a contrast between academic and non-academic sources for our analyses.

4.3 Preprocessing steps

All corpora were preprocessed in the same way: we removed punctuation, numbers, stop-words and non-English words, did case folding and lemmatization using spaCy.[3]

The resulting corpus of psychology abstracts comprises 73,788,954 tokens from 825,628 texts. The general domain corpus has 253,597 texts with 237,205,654 tokens in total.

5 Representation of concepts

We manually associate each concept with a list of most morphologically and semantically related words. For our frequency analysis we sum the corresponding token frequencies.[4] We only consider tokens that occurred at least 50 times in each corpus. The final representation of concepts is as follows:

‘Addiction’: *addict*, *addiction*

‘Bullying’: *bully*, *bullying*

‘Trauma’: *trauma*, *traumatic*, *traumatize*

‘Harassment’: *harass*, *harassment*

‘Prejudice’: *prejudice*

[3] https://spacy.io/. spaCy uses a pre-trained multi-task CNN-based model that takes into account part-of-speech information (i.e. adjective *addicted* will not be transformed into *addict*).

[4] As we mentioned above, the corpus contains lemmata only.

In order to obtain concept vector representations, we follow the DISTRIBUTED DICTIONARY REPRESENTATIONS approach proposed in Garten et al. (2018) which is similar to Mendelsohn et al. (2020). More specifically, we represent each concept as a mean vector of the corresponding word vector representations (e.g., ‘addiction’ would be an average of vector representations of *addict* and *addition*). Unlike Mendelsohn et al. (2020) we do not assign frequency-based weights to tokens.

6 Experiments

6.1 Frequency-based analysis

For each of the five concepts we first evaluate their (unigram) frequency distribution over time. We evaluate relative frequencies by normalizing the raw counts by the total number of tokens in each year.[5]

As Figure 2.2 demonstrates, in the psychology domain all concepts demonstrate relative increase in frequency: ‘trauma’ exhibits the steepest slope, ‘bullying’ gradually raises since the 1990s, and ‘harassment’ has its peak in the mid-1990s. ‘Addiction’ and ‘prejudice’ present the lowest changes in relative frequency. The results obtained on CoCA/COhA (Figure 2.3) are more unsteady and labile: ‘trauma’ rises over time but much less rapidly compared to the psychology literature, relative frequencies of ‘addiction’ and ‘bullying’ increase over time. ‘Harassment’ also demonstrates the highest usage in the early 1990s while ‘prejudice’ slightly declines. Does the increase in the frequency of ‘trauma’ imply that it has broadened over time, i.e. its usages expanded to new contexts, especially in psychology literature? On the other hand, ‘trauma’ exhibits the highest usage among the five concepts in psychology literature, so “the law of conformity” (Hamilton et al. 2016b) would predict that it should change slower. ‘Harassment’ presents the lowest raw frequencies throughout most time periods but has risen in the mid-nineties. Would this imply that ‘harassment’ changed its meanings faster and achieved the highest breadth in the nineties?

In the next section, we adapt two diachronic variations of word2vec (Mikolov, Sutskever, et al. 2013) to quantify semantic change over time. We first train a type-based model conceptually similar to the one proposed in Mendelsohn et al. (2020). We use the type-level embeddings to obtain token-level (sentence-specific) representations which are further utilized to measure semantic breadth in each epoch.

[5] We also applied a moving average smoothing with window size of 1, i.e. $f_{1972} = (f_{1971} + f_{1972} + f_{1973})/3$.

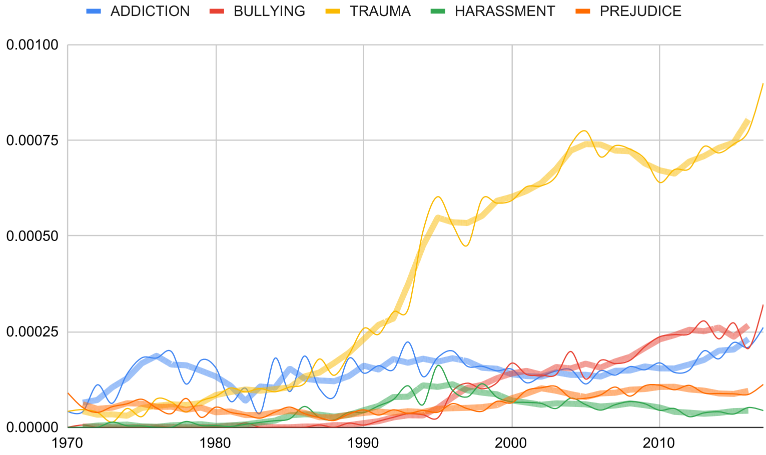


Figure 2.2: Relative concept frequencies based on abstracts from psychology journals. Bold lines correspond to moving average smoothing (window=1).

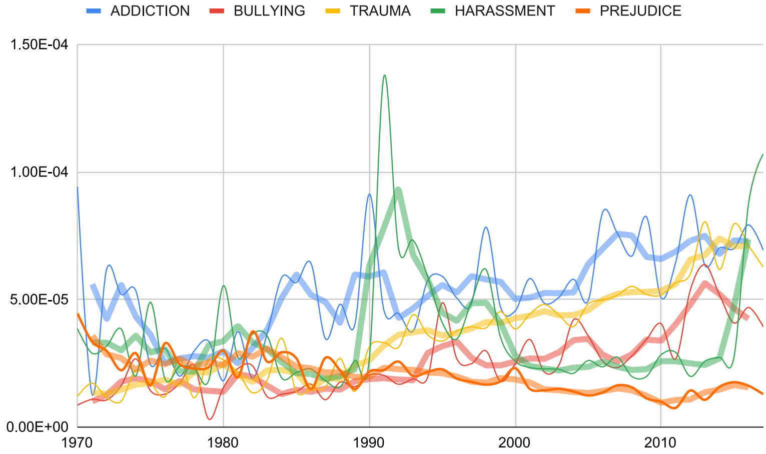


Figure 2.3: Relative concept frequencies based on general domain corpus. Bold lines correspond to moving average smoothing (window=1).

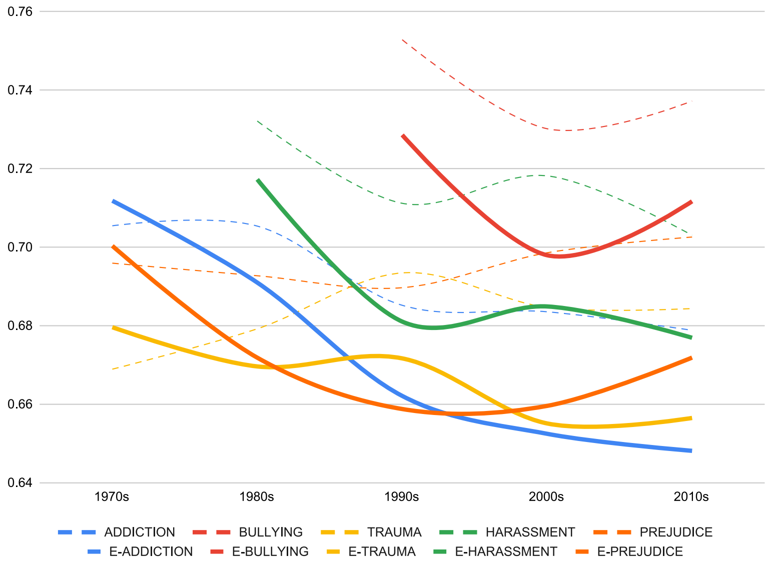


Figure 2.4: Mean cosine similarities (polynomial smoothing) over five decades (psychology abstracts corpus). Bold and dashed lines correspond to epoch-specific (e-*) and global (static) embeddings, respectively.

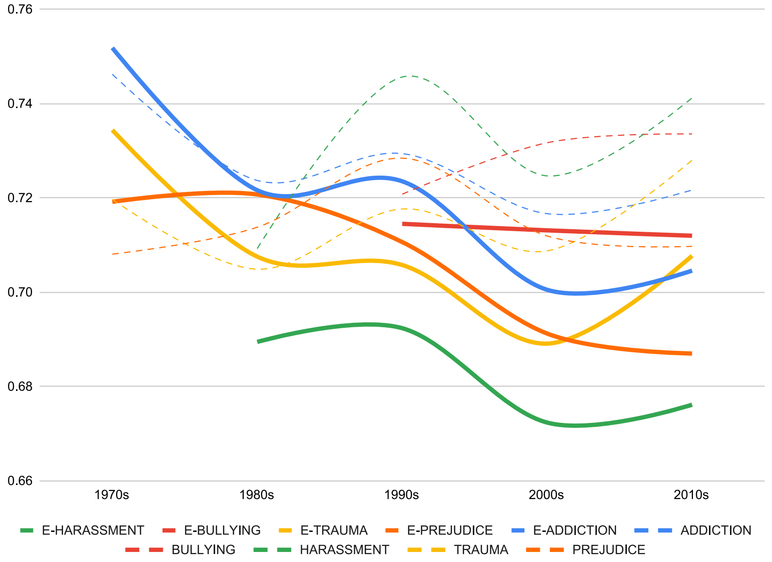


Figure 2.5: Mean cosine similarities (polynomial smoothing) over five decades (general domain corpus). Bold and dashed lines correspond to epoch-specific (e-*) and global (static) embeddings, respectively.

We then take a closer look at the type-level epoch-specific embeddings to study *how exactly* concepts changed. Such models have previously shown their utility at capturing semantic changes over time (Tahmasebi et al. 2018, Kutuzov et al. 2018) and do not require vector space alignment (which, as has been previously shown, leads to noise; Dubossarsky et al. 2019).

6.2 Diachronic word2vec

We first train a *type-level* word2vec skip-gram model.[6] In terms of hyper-parameter setting we follow that of Mendelsohn et al. (2020). Since we mainly focus on semantic changes, we set the context window size to 10 to better capture semantics and associations (Agirre et al. 2009). We also do not consider tokens that occur less than 5 times over the whole corpus. We train the model on the whole corpus for 10 iterations (obtaining *global*–static embeddings). We then use the global embeddings to initialize epoch-specific models that we continue training on each epoch's data independently for another 10 iterations. We split time periods by decades.[7]

6.2.1 Token-level embeddings

In order to obtain token-level embeddings, the resulting (global and epoch-specific) embeddings are then contextualized for each decade starting the 1970s and finishing the 2010s. This part of experiments is based on the method proposed by Sagi et al. (2009) except that we use the word2vec model (Mikolov, Sutskever, et al. 2013) rather than LSA (Landauer et al. 2013) (therefore, we refer to it as "neural parameterization of Sagi et al.'s model").

More specifically, in order to get sentence-specific vector representations for each concept in a certain decade, we randomly sample a number of its sentential occurrences[8] from the respective period, then extract contextual tokens found within the pre-set window size.[9] The final sentence-specific representation is a bag-of-words, i.e. it is an average over corresponding token representations. Following Sagi et al. (2009), in order to estimate the semantic breadth of a word,

[6] Using https://radimrehurek.com/gensim/.

[7] Due to an insufficient amount of data for earlier time periods, we train the models only on the following time frames: 1980–1989, 1990–1999, 2000–2009, 2010–2017.

[8] We set the number to 50. We use all sentential instances if the concept occurs less than 50 times during the epoch (having 20 as a minimum)

[9] We set the window size to be 3, 7, 9 tokens at each side and found that 9 provides smoother results, so we used this setting throughout.

we evaluate pair-wise cosine similarities across all the sentence-specific representations. To reduce any biases, we repeat the above sentence sampling process 10 times. The final mean values for cosine similarities for both types of models, global and epoch-specific, in the psychology and general domains are presented in Figures 2.4 and 2.5. The figures also illustrate that epoch-specific embeddings (marked as bold) provide more robust results, and we will mainly rely on them in our study.

The five concepts behave differently over time. For instance, ‘trauma’, although becoming frequently used in the psychology corpus, has only broadened its meaning slightly and has stayed quite a “broad” concept. In CoCA/CoHA *trauma* does not appear much before the 1990s.[10] Figure 2.5 presents two slopes; the first one can be possibly explained by the difference in its frequency distribution in CoCA and COhA, while the second one is due to its breadth changes. The notion of ‘harassment’, on the other hand, has the steepest slope between the 1980s and the 1990s, and then it stabilizes in its contextual usages. The highest contextual similarity in the 1980s can be partially attributed to relatively few usage instances in psychology corpus during this period. In CoCA/CoHA, frequency of ‘harassment’ has a drastic leap in the 1990s but, as Figure 2.5 shows; it does not affect its breadth when compared to the 1980s, although it becomes broader in the the 2000s (its usage frequency also decreases). The concept of ‘bullying’ has been constantly increasing in its relative usage frequency in the psychology literature, although its semantics presents a more complex pattern: it broadened from the 1990s to the 2000s, and then narrowed in the 2010s. In CoCA/CoHA the usage of ‘bullying’ was more stable and did not significantly change in frequency and semantics. Similarly, ‘addiction’ stayed within the same frequency range after the 1990s (although being much less frequent in the 1970s and the 1980s), and its breadth slightly increased since then. In the psychology domain its semantic breadth changes are more drastic: ‘addiction’ has been gradually becoming broader since the 1970s due to its expansion to new behavior types. Finally, ‘prejudice’, the concept that was not widely used before the 2000s in both corpora,[11] behaves differently in the general and psychology domains: in psychology abstracts it narrows down in the 2010s while in CoCA/CoHA it continues to expand its meaning. The results support the findings obtained for the LSA-based model in Vylomova et al. (2019). The next part of the chapter investigates how exactly the meanings changed.

[10] I.e. it is much less represented in CoHA.

[11] It appears less than 100 times a year before early 2000s.

6.2.2 Type-level embeddings

We now use the obtained epoch-specific *type-level* embeddings to run a detailed study of concept change.

Following Hamilton et al. (2016b), we consider two metrics to evaluate semantic changes over time:

1. *Semantic displacement*, which shows to what extent a concept has semantically changed during a certain time period. This is quantified as cosine distance between the word embeddings from the corresponding time periods, i.e. cos-dist($\mathbf{w}^t, \mathbf{w}^{t+\delta}$).

   Figure 2.6 shows the results of the semantic displacement evaluation and confirms our observations made earlier using the model from Sagi et al. (2009). Concepts such as 'trauma', 'bullying', 'prejudice' change similarly in the psychology and general domain corpora. The largest gaps are observed in the case of 'addiction'.

2. *Pair-wise similarity time-series*, which is quantified as

   $$s^{(t)}(w_i, w_j) = \text{cos-sim}(\mathbf{w}_i^t, \mathbf{w}_j^t)$$

   and measures how cosine similarity between words w_i and w_j changes over time period $(t; t+\delta)$. For each concept we first constructed a list of words which the concept most often co-occurred with within each time period. Then we calculated cosine similarity between the concept and every word from the list for each decade. We will now discuss changes in each concept individually.

6.2.2.1 'Trauma'

As Figure 2.7 illustrates, 'trauma' has undergone more significant meaning changes in the psychology literature than in CoCA/CoHA where it preserves most associations since the 1990s. More specifically, in the psychology corpus, we observe a clear shift from *physical* to *psychological.* Although its relatedness to *severe* is still more prevalent than *mild*, they both increase their similarity to 'trauma' over time. In both corpora, 'trauma' started moving away from *childhood* in the 2000s.

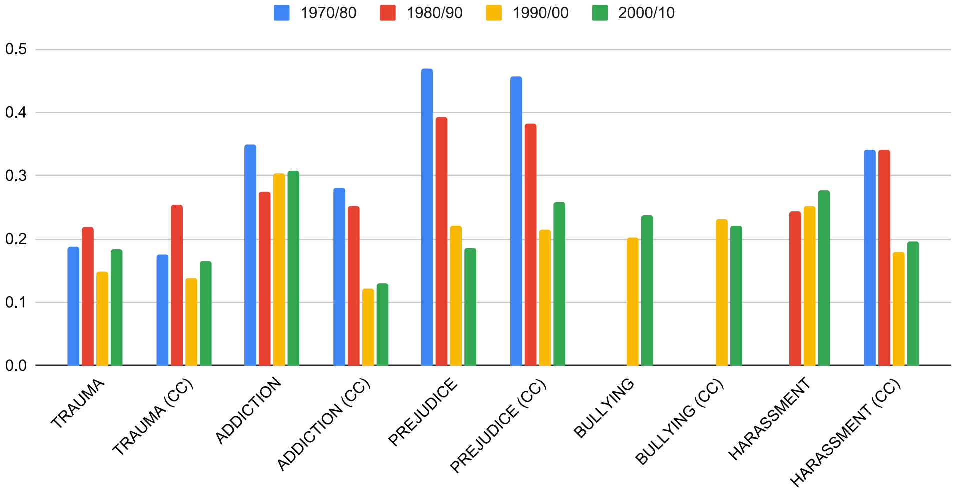
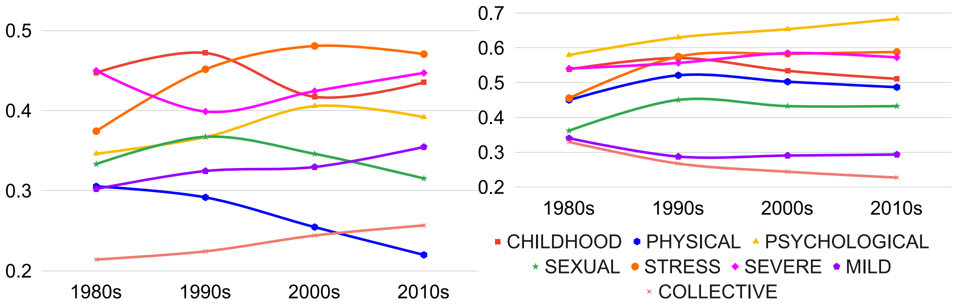


Figure 2.6: Cosine distances between decades in the psychology and CoCA/CoHA (CC) domains



(a) Psychology

(b) General Domain (CoCA+CoHA)

Figure 2.7: ‘Trauma’. Cosine similarities over four decades

Table 2.1 lists its top nearest neighbors in both corpora: ‘trauma’ stays strongly associated with ‘PTSD’. In the general domain it is associated with *horrific* and *suffer*, and its relatedness to the latter increases over time. During the 1990s–2000s ‘trauma’ becomes more *emotional* and *psychological*, which is in line with Haslam & McGrath (2020)’s findings that show changes in the relative frequency of trauma-related concepts in the massive Google Books corpus from 1960 to 2008. They found that during the 1990s the term *psychological trauma* rose most steeply.

Table 2.1: ‘Trauma’. Top-10 nearest neighbors

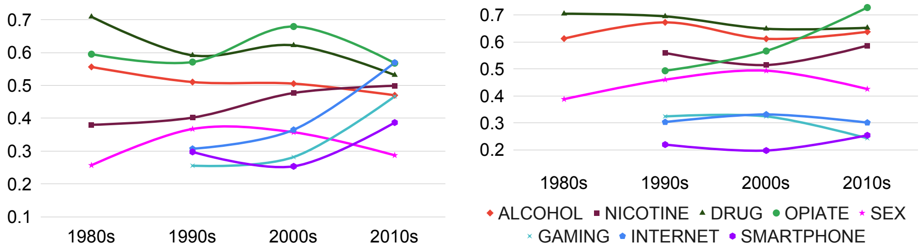
Psychology			
1980s	1990s	2000s	2010s
humbling	posttraumatic	ptsd	posttraumatic
posttraumatic	ptsd	posttraumatic	ptsd
retraumatization	survivor	traumatization	traumatization
traumatized	posttrauma	traumatized	aftermath
traumatizing	retraumatization	traumatizing	dissociative
traumatogenic	injury	desnos	peritraumatic
terrifying	traumatized	torture	traumatized
debility	atrocity	survivor	traumatically
traumatise	traumatization	dissociative	posttrauma
survivor	dissociative	posttrauma	atraumatic
traumatization	sequelae	flashback	traumatizing
unassimilable	ptsdlike	nontraumatize	pts
torture	ptds	retraumatization	refugee
traumatised	peritrauma	peritraumatic	mtbi
hypnoanalysis	desnos	nontrauma	telecommunicator
traumatolytic	psychotraumatic	lifethreat	ptss
keilson	torture	holocaust	sequelae
flashback	reexperience	nontraumatic	flashback
psychotraumatic	lasc	traumatise	postraumatic
hypnoid	traumatologist	ptes	desnos
CoCA/CoHA			
1980s	1990s	2000s	2010s
salomo	ptsd	ptsd	ptsd
posttraumatic	spiegle	posttraumatic	psychological
hyperarousal	psychological	traumatization	boehnlein
traumatization	emotional	psychological	posttraumatic
reliving	posttraumatic	emotional	hyperarousal
traumatized	horrific	horrific	suffer
indentify	psychosis	suffer	traumatization
louxes	traumatization	traumatizing	horrific
clinginess	suffer	experiencing	emotional
emotional	victim	hyperarousal	experiencing
przekop	disorder	spiegle	spiegle
brayme	sexualizing	przekop	injury
experiencing	hyperarousal	disorder	victim
ptsd	brayme	victim	traumatizing
boehnlein	abuse	painful	disorder
rohrbacher	syndrome	brayme	traumatically
spiegle	therapist	hospitalize	csf2
traumatizing	cope	severe	scurfield
csf2	boehnlein	psychiatric	przekop
traumatically	gavigan	tbi	yancosek

Table 2.2: ‘Addiction’. Top-10 nearest neighbors

Psychology			
1980s	1990s	2000s	2010s
heroin	addicted	addicted	addictive
addicted	addictive	opiate	addicted
narcotic	abuser	abuser	dependence
methadone	heroin	heroin	heroin
nonopiate	substance	addictive	mmt
illicit	dependence	drug	craving
nonaddiction	alcoholic	dependence	internet
drug	drug	substance	opiate
opiate	opiate	methadone	opioid
alcoholism	methadone	abstinence	drug
polysubstance	cocaine	cocaine	abuser
detoxification	alcoholism	detoxify	cybersex
nonaddicte	gambler	illicit	substance
alcohol	crack	detoxification	crave
abuser	abuse	abuse	problematic
detox	nonaddicte	crave	yfas
mmt	coaddict	craving	detoxification
cocaine	alcohol	abstinent	igd
polydrug	detoxification	opioid	abstinent
nonnarcotic	mmt	alcoholism	gaming
CoCA/CoHA			
1980s	1990s	2000s	2010s
drug	addicted	addicted	addicted
heroin	heroin	heroin	heroin
pcp	drug	addictive	opiate
abuser	abuser	alcoholism	opioid
methadone	alcoholic	meth	methadone
cocaine	cocaine	cocaine	alcoholism
marijuana	alcoholism	alcoholic	rehab
amphetamine	methadone	abuser	addictive
addictive	alcohol	rehab	alcoholic
alcohol	addictive	drug	suboxone
opioid	rehab	oxycontin	drug
alcoholic	oxycontin	methamphetamine	painkiller
quashen	marijuana	waismann	quashen
cannabis	abuse	medicate	overdose
opiod	henningfield	methadone	alcohol
alcoholism	buprenorphine	alcohol	vivitrol
methamphetamine	12step	quashen	acamprosate
mdma	quashen	buprenorphine	relapse
mcshin	relapse	12step	sober
addicted	addicting	relapse	oxycontin

6.2.2.2 ‘Addiction’

‘Addiction’ demonstrates a remarkable shift in the psychology literature from a substance-related concept in the 1980s to a behavior-related concept in the 2010s, but this pattern is less evident in CoCA/CoHA (see Figure 2.8 and Table 2.2). More specifically, we observe that the concept moved away from ‘narcotic’-related meanings towards *gaming*, *Internet*, *cybersex*, and *smartphone*. The findings confirm earlier observations done by Vylomova et al. (2019) who used the diachronic language model from Hamilton et al. (2016b). In psychology literature, such conceptual expansion of ‘addiction’ had prompted and induced adaptation of a range of psychosocial treatments to be used to treat gambling, internet, and sexual addictions (Yau & Potenza 2015).



(a) Psychology

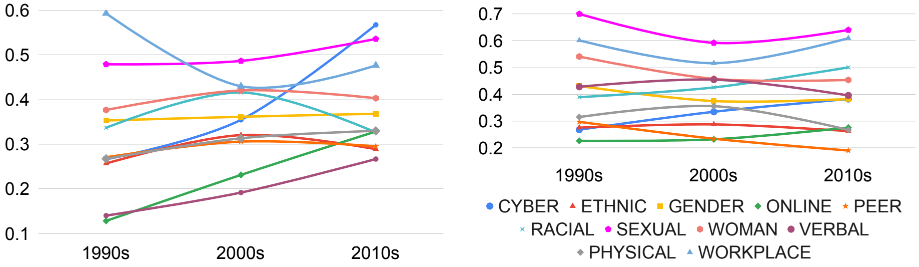
(b) General Domain (CoCA+CoHA)

Figure 2.8: ‘Addiction’. Cosine similarities over four decades.

In the general domain corpora, initial associations of ‘addiction’ are more stable over time, and the similarity to *opiate* even increases during the last two decades. In both domains, ‘addiction’ becomes less associated with *abuse* and *abuser*: the similarity drops by 0.1–0.15 since the 1990s and 2000s.

6.2.2.3 ‘Harassment’

In both corpora, usage of ‘harassment’ increases in the 1990s, and the 1980s do not contain enough instances to obtain reliable embeddings. As Figure 2.9 shows, ‘harassment’ is highly related to *sexual* in both domains. In the psychology literature ‘harassment’ moves away from *workplace* towards *online* and *cyber* (increasing its relatedness to ‘bullying’). In the general domain there are fewer marked changes across decades. The relationship to *online* and *cyber* is weaker than in the psychology corpus and, in contrast to that corpus, ‘harassment’ is more associated with *verbal* than *physical*.[12] These findings point to similarities across the corpora, but we observe a more rapidly growing preoccupation of psychology with digitally mediated forms of harassment.



(a) Psychology

(b) General Domain (CoCA+CoHA)

Figure 2.9: 'Harassment'. Cosine similarities over three decades

By looking at the nearest neighbors space shown in Table 2.3, we additionally notice substantial differences in the two domains: 'harassment' in psychology preserves its emphasis on *victimization*, the act or process of singling someone out for cruel or unfair treatment, typically through physical or emotional abuse,[13] while increases that of *perpetration*. During the 2000s–2010s it reduced its relatedness to *violence*. The general domain treats the concept of 'harassment' in a somewhat more legalistic frame, as a form of *misconduct* that is tightly associated with *allegation*, *complaint*, *accusation*, and *abuse*.

6.2.2.4 'Bullying'

Similarly, 'bullying' is markedly more victim-related in the psychology domain, having both *victimization* and *perpetration* among its top nearest neighbors in the 2000s–2010s. We additionally observe an increase in its association with 'harassment'. As shown in Figure 2.10, it becomes more associated with *workplace* while its similarity to *school* and *child* rises less steeply, consistent with its expansion into the adult realm. Similar to other concepts, we observe that bullying has expanded to cyberspace. Interestingly, its association with *cyber* accelerates upwards faster than the other concepts. As Haslam (2016) notes, referring to indirect, digitally mediated forms of aggression as "cyber-bullying" is a paradigm case of horizontal concept creep.

[12] This is probably due to 'physical' being the default characteristic of 'harassment' and usually is not explicitly marked.

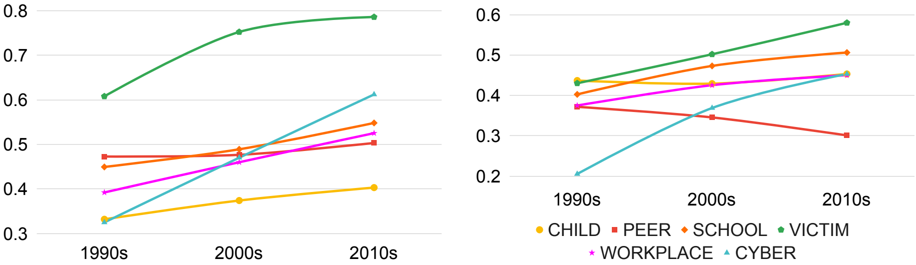
[13] The definition provided in https://dictionary.apa.org/victimization.

Table 2.3: ‘Harassment’. Top-10 nearest neighbors

	Psychology	
1990s	2000s	2010s
harasser	harasser	harasser
workplace	victimization	victimization
contrapower	contrapower	cyber
neosh	uncivil	assault
harassing	assault	bullying
uncivil	victim	perpetration
coercion	bullying	victimize
unprofessional	perpetrator	bully
assault	lsh	victim
nonharasse	victimize	perpetrate
gutek	violence	bystander
rape	perpetration	homophobic
sexualized	perpetrate	cyberbullying
nonheterosexist	rape	incivility
sexual	bully	insinuation
coercive	lgbts	heterosexist
employee	harassing	contrapower
perpetrator	cyberstalking	cyberbullye
incident	socialsexual	sexual
intragender	cyberbullye	violence
	CoCA/CoHA	
1990s	2000s	2010s
harasser	complaint	harasser
sexual	accuse	allegation
misconduct	complain	assault
complaint	intimidate	sexual
allege	assault	accusation
eeoc	intimidation	allege
accuser	allegation	complaint
sexually	harasser	workplace
allegation	misconduct	accuser
accuse	abuse	accuse
accusation	abusive	misconduct
lawsuit	threaten	rape
assault	accusation	intimidation
discrimination	renaye	lawsuit
incident	sue	harassing
abusive	discrimination	alleged
workplace	allege	defamation
abuse	sexual	abuse
rape	sexually	bullying
intimidation	mutziger	mistreatment

Table 2.4: ‘Bullying’. Top-10 nearest neighbors

	Psychology	
1990s	2000s	2010s
victimisation	bullyvictim	cyberbullying
victimise	victimization	cyberbullye
cyberbullying	cyberbullying	victimization
olweus	victimisation	cyber
bullyvictim	cyberbullye	victimisation
antibullying	victimise	cyberbully
victim	olweus	perpetration
namecalling	notinvolve	antibullying
cyberbullie	victim	victimize
victimize	antibullye	victim
provictim	antibullying	harassment
cyberbullye	nonbullye	olweus
antibullye	cybervictim	antibullye
cyberbully	victimize	cyberbullie
bystanding	dipc	bystander
ringleader	cyberbullie	bystanding
bullycategorie	bullycategorie	cybervictimization
notinvolve	cyberbully	bullyvictim
nonbullye	bystander	defending
victimization	selfdestruction	kiva
	CoCA/CoHA	
1990s	2000s	2010s
aggression	bullied	cyberbullie
olweus	olweus	cyberbullying
schoolyard	coloroso	bullied
taunt	taunt	abuse
humiliate	harassment	olweus
behavior	abuser	cyberbullye
intimidate	cyberbullie	harassment
punish	himel	mutziger
abusive	intimidate	vanheest
harass	behavior	cyberbully
abuse	abuse	mehus
intimidation	13er	intimidation
aggressive	montooth	kiongozi
prosocial	milvin	nishina
taunting	nishina	taunt
13er	marrinson	insensitively
angry	aggression	zirpola
bullied	namie	sharaud
skutch	weinsheimer	fifthgrade
aggressor	vanheest	harasser



(a) Psychology

(b) General Domain (CoCA+CoHA)

Figure 2.10: ‘Bullying’. Cosine similarities over three decades

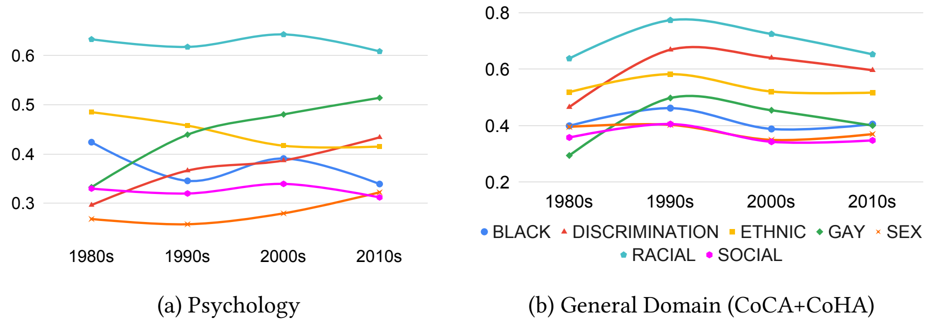
In the psychology literature, ‘bullying’ is also strongly intertwined with ‘harassment’, and both are linked to the notion of *victimization*. Arguably, this strong focus on victimization in the psychological literature, also evident in the concept of ‘harassment’, represents a preoccupation with the harm caused by bullying. The results obtained on CoCA/CoHA appear to be less congruent and more noisy, and emphasize the behaviors involved in bullying rather than the harmful impact they have on their targets. Still, it is clear that ‘bullying’ becomes more closely related to *abuse* over time in that corpus but less related to *aggression*.

6.2.2.5 ‘Prejudice’

In both corpora, but especially in psychology, ‘prejudice’ is highly associated with *racial* or *racism*, both of which are also among its nearest neighbors during all decades (see Table 2.5). In the psychology corpus, the similarity is relatively stable while in CoCA/CoHA it reduces over time. The association of ‘prejudice’ with *ethnic* and *ethnicity*, on the other hand, drops in both corpora. Dynamics of similarity with *discrimination* presents differences: it decreases in CoCA/CoHA while it rises (along with similarity to *anti-discrimination*) in psychology. The same pattern can be observed for *gay*. Interestingly, in the psychology corpus *anti-gay* and *pro-gay* are among the nearest neighbors and the similarity with *both* of them increases over time, indicating a rising attention to anti-gay prejudice within psychology over time that is not seen in the general domain. This represents a “horizontal” expansion of ‘prejudice’ in psychology beyond its earlier exclusive focus on racial animosity.[14] Analysis of nearest neighbors shows that in both domains the associations between ‘prejudice’ and *stereotyping*, *bigotry* and

[14] Among 200 nearest neighbors in each decade, the number of “anti-” and “pro-” terms is higher in psychology than in CoCA/CoHA.

belief are among the strongest and most stable over time. In the psychology literature ‘prejudice’ increases its similarity to *micro-assault* and *micro-insult* over the last decade. The growing relatedness to these forms of “micro-aggression” (Lilienfeld 2017) supports the claim that ‘prejudice’ has crept “vertically” to encompass increasingly subtle phenomena.



(a) Psychology

(b) General Domain (CoCA+CoHA)

Figure 2.11: ‘Prejudice’. Cosine similarities over four decades.

7 Conclusion

The findings of our analyses illuminate and add nuance to our understanding of concept creep within academic psychology and general domain corpora. The diachronic analysis reveals a trend for our sample of harm-related concepts to undergo semantic broadening from the 1970s to the 2010s, although the trajectories of particular concepts have been neither consistent nor linear. Since the 1990s, for example, ‘addiction’, ‘bullying’ and ‘harassment’ have broadened, as the theory of concept creep would suggest, but the breadth of ‘trauma’ and ‘prejudice’ have been relatively static. The changes are more evident in psychology literature compared to CoCA/CoHA. The analysis of semantic displacement points to a more consistent diachronic pattern: the majority of concepts changed most substantially from the 1980s to the 1990s and changed progressively less thereafter. This finding implies that societal and cultural changes occurring in the final two decades of the 20th century are likely to be especially critical for understanding concept creep. Finally, the analysis of pairwise similarities demonstrated changing patterns of co-occurrence for each concept that clarified how its meanings have shifted and expanded over four decades. During this period some concepts have acquired entirely new associations (e.g., *cyber-harassment*), some have added new semantic domains (e.g., ‘addiction’ incorporating non-ingestive

Table 2.5: ‘Prejudice’. Top-10 nearest neighbors

Psychology			
1980s	1990s	2000s	2010s
prejudiced	prejudiced	prejudiced	prejudiced
ethnocentrism	antiblack	intergroup	intergroup
xenophobia	antiforeigner	stereotyping	blatant
racial	stereotyping	blatant	stereotyping
prejudicial	stereotype	outgroup	outgroup
racism	compunction	derogated	rwa
racist	prejudicial	sdo	sdo
neuroessentialism	antigay	justif.suppression	authoritarianism
postcivil	racism	racism	derogated
ethnic	ethnopolitical	racist	antigay
ethnocentric	antifat	microinsult	justif.suppression
justif.suppression	tropp	minoritygroup	homophobia
sexblindness	antiatheist	majoritygroup	ideology
sdo	antihomosexual	antigay	rightwing
transprejudice	justif.suppression	antihomosexual	minoritygroup
intelligentsia	oldfashioned	antiblack	microassault
antiblack	neosexist	prejudicial	tropp
favoritism	intergroup	microinvalidation	antiforeigner
microinsult	multiculturalist	ingroup	microinsult
eugenics	problack	nonprejudicial	progay

CoCA/CoHA			
1980s	1990s	2000s	2010s
bigotry	racism	racism	bigotry
stereotyping	bias	bigotry	racism
racism	racial	racial	discrimination
racialist	bigotry	stereotype	stereotype
halfheartedness	discrimination	discrimination	racial
elitism	stereotype	injustice	ignorance
racial	prejudiced	racist	racist
belief	ignorance	colorism	belief
outsiderness	racist	hatred	oppression
uncomplicatedly	stereotyping	bias	bias
delegitimate	hatred	homophobia	classism
ridiculing	oppression	bigoted	misogyny
muddleheaded	injustice	belief	sexism
ethnocentrism	bigot	nonwhite	notion
factionalize	animosity	hostility	hatred
animus	homophobia	religion	discriminate
fact	bigoted	ignorance	denesh
biologism	sexism	speciesism	colorism
snideness	gender	semitism	prejudiced
multiculturalist	distrust	heterosexism	ridiculing

behaviors such as gaming and smartphone use), others have shifted emphasis (e.g., ‘trauma’ becoming associated less with physical injury and more with psychological stress), and yet others have come to refer to less severe phenomena (e.g., ‘prejudice’ becoming associated with so-called micro-aggressions). Collectively, these findings support the presence of both horizontal and vertical concept creep as proposed by Haslam (2016).

The results of the present analyses are in some respects preliminary. From a methodological standpoint, future research will need to optimize the analytic parameters employed in the approaches examined in this research and evaluate whether findings derived from these approaches converge with those using other methods for assessing semantic change. Methods must also be developed to examine horizontal and vertical concept creep separately. The methods used in the present research emphasize “horizontal” changes in the range of semantic contexts in which a concept appears, and are less capable of capturing how meanings may shift “vertically” to encompass less severe phenomena. The latter can only be inferred indirectly when concepts referring to such subtler phenomena become increasingly near semantic neighbors of the target concept.

Substantively, our findings should be replicated with additional hypothetically creeping concepts, such as ‘mental illness’ and ‘safety’. The extent to which expansionary semantic changes are specific to harm-related concepts rather than generalized must also be studied systematically. There is scope for more focused and finely detailed analyses of semantic shifts in single concepts. Indeed, our approach offers a versatile methodology for evaluating the nature, timing, and nearest-neighbor subtleties of such shifts. Ideally, future work will explore concept creep in corpora representing other scholarly disciplines and other languages. A more fundamental challenge is to uncover the cultural factors that contribute to the semantic inflation of harm-related concepts, and to understand its societal implications.

Abbreviations

ACL	Association for Computational Linguistics
LSA	latent semantic analysis
PTSD	post-traumatic stress disorder
SVD	singular value decomposition
TF-IDF	term frequency - inverse document frequency

References

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca & Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL 2009*, 19–27. Boulder: ACL.

Antoniak, Maria & David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the ACL* 6. 107–119.

Belinkov, Yonatan & James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the ACL* 7. 49–72.

Bender, Emily M. & Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL 2020*, 5185–5198. Online: ACL. DOI: 10.18653/v1/2020.acl-main.463.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Bréal, Michel. 1897. *Essai de sémantique*. Paris: Hachette.

Davies, Mark. 2008. *The corpus of contemporary American English (COCA): 560 million words, 1990-present.*

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora* 7(2). 121–157.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of ACL 2019*, 457–470. Florence: ACL. DOI: 10.18653/v1/P19-1044.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer & Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of NetWordS 2015* (CEUR Workshop Proceedings 1347), 66–70. Pisa.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Fabiano, Fabian & Nick Haslam. 2020. Diagnostic inflation in the DSM: A meta-analysis of changes in the stringency of psychiatric diagnosis from DSM-III to DSM-5. *Clinical Psychology Review* 80. 101889. DOI: 10.1016/j.cpr.2020.101889.

Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis*, 1–32. Reprinted in: Palmer, F. R. (ed.) .1968. Selected Papers of J. R. Firth 1952-59. 168-205. London: Longmans. Hoboken: Basil Blackwell.

Garten, Justin, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch & Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods* 50(1). 344–361.

Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat: Een onderzoek naar kleding-en voetbaltermen.* Amsterdam: Meertens-Instituut.

Gladkova, Anna, Aleksandr Drozd & Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL student research workshop*, 8–15. San Diego.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of EMNLP 2016*, 2116–2121. Austin: ACL. DOI: 10.18653/v1/D16-1229.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2-3). 146–162.

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

Haslam, Nick, Brodie C. Dakin, Fabian Fabiano, Melanie J. McGrath, Joshua Rhee, Ekaterina Vylomova, Morgan Weaving & Melissa A. Wheeler. 2020. Harm inflation: Making sense of concept creep. *European Review of Social Psychology* 31(1). 254–286.

Haslam, Nick & Melanie J. McGrath. 2020. The concept creep of trauma. *Social Research: An International Quarterly* 87(3). 509–531.

Jurgens, David & Keith Stevens. 2009. Event detection in blogs using temporal random indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, 9–16. ACL.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore: ACL. DOI: 10.3115/v1/W14-2517.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*, 1384–1397. Santa Fe: ACL.

Landauer, Thomas K., Danielle S. McNamara, Simon Dennis & Walter Kintsch. 2013. *Handbook of latent semantic analysis*. Hove: Psychology Press.

Lehrer, Adrienne. 1985. The influence of semantic fields on semantic change. In *Historical semantics – historical word-formation* (Trends in Linguistics. Studies and Monographs [TiLSM] 29), 283–296. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110850178.283.

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

Lilienfeld, Scott O. 2017. Microaggressions: Strong claims, inadequate evidence. *Perspectives on Psychological Science* 12(1). 138–169.

Lukianoff, Greg & Jonathan Haidt. 2019. *The coddling of the American mind: How good intentions and bad ideas are setting up a generation for failure*. London: Penguin Books.

McGrath, Melanie J., Kathryn Randall-Dzerdz, Melissa A. Wheeler, Sean C. Murphy & Nick Haslam. 2019. Concept creepers: Individual differences in harm-related concepts and their correlates. *Personality and Individual Differences* 147. 79–84.

Mendelsohn, Julia, Yulia Tsvetkov & Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *arXiv preprint arXiv:2003.03014*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in neural information processing systems*, 3111–3119. Red Hook, NY: Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL 2013*, 746–751. Atlanta: ACL.

Pinker, Steven. 2018. *Enlightenment now: The case for reason, science, humanism, and progress*. London: Penguin.

Rogers, Anna, Olga Kovaleva & Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 104–111. Athens: ACL. https://www.aclweb.org/anthology/W09-0214.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of SemEval 2020*, 1–23. Barcelona: ACL. https://www.aclweb.org/anthology/2020.semeval-1.1.

Stern, Gustaf. 1931. *Meaning and change of meaning; with special reference to the English language*. Gothenburg: Wettergren & Kerbers.

Sturtevant, Edgar Howard. 1917. *Linguistic change: An introduction to the historical study of language*. Vol. 60. Chicago, IL: University of Chicago Press.

Sunstein, Cass R. 2018. *The power of the normal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3239204.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278*.

Vylomova, Ekaterina, Sean Murphy & Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 29–34. Florence: ACL. DOI: 10.18653/v1/W19-4704.

Vylomova, Ekaterina, Laura Rimell, Trevor Cohn & 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 ACL 2016*, vol. 1: Long Papers, 1671–1682. Berlin.

Weaver, Warren. 1955. Translation. *Machine translation of languages* 14(15-23). 10.

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

Winter, Bodo, Graham Thompson & Matthias Urban. 2014. Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In *Proceedings of EVOLANG 2014*, 353–360. Utrecht: World Scientific.

Xu, Yang & Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual meeting of the Cognitive Science Society, CogSci 2015*. Pasadena.

Yau, Ms Yvonne HC & Marc N. Potenza. 2015. Gambling disorder and other behavioral addictions: Recognition and treatment. *Harvard Review of Psychiatry* 23(2). 134.