Evaluation of Semantic Change of Harm-Related Concepts in Psychology

Ekaterina Vylomova Sean Murphy Nick Haslam

School of Psychological Studies, University of Melbourne, Melbourne, Australia
{vylomovae,nhaslam}@unimelb.edu.au seanchrismurphy@gmail.com

Abstract

The paper focuses on diachronic evaluation of semantic changes of harm-related concepts in psychology. More specifically, we investigate a hypothesis that certain concepts such as "addiction", "bullying", "harassment", "prejudice", and "trauma" became broader during the last four decades. We evaluate semantic changes using two models: an LSA-based model from Sagi et al. (2009) and a diachronic adaptation of word2vec from Hamilton et al. (2016), that are trained on a large corpus of journal abstracts covering the period of 1980–2019. Several concepts showed evidence of broadening. "Addiction" moved from physiological dependency on a substance to include psychological dependency on gaming and the Internet. Similarly, "harassment" and "trauma" shifted towards more psychological meanings. On the other hand, "bullying" has transformed into a more victim-related concept and expanded to new areas such as workplaces.

1 Introduction

During the last decade the area of diachronic language modelling has witnessed substantial progress and development. This technical development enables enhanced understanding of pressing issues in social science disciplines. In this paper, we focus on diachronic change in the semantics of harm-related concepts within psychology. In particular, we test a "concept creep" hypothesis proposed by Haslam (2016). The hypothesis states that during the last half century many concepts related to harm have broadened their meanings. In order to test the hypothesis, we utilize two diachronic models: a count-based approach from Sagi et al. (2009), and a prediction-based approach from Hamilton et al. (2016). In both cases, we estimate the breadth of a concept as its average cosine similarity, i.e. the lower the similarity between concepts vector representations, the broader the concept's meaning. We additionally investigate how exactly the meanings have changed.

2 The Notion of Concept Creep

Haslam (2016) put forward the notion of concept creep to describe an apparent expansion in the meanings of several fundamental psychological concepts. He presented a series of case studies in which psychological researchers and theorists expanded the sense of these concepts by loosening definitions to include milder instances ("vertical creep") or by extending definitions to encompass cognate phenomena ("horizontal creep"). More example, the concept of "mental disorder" has progressively broadened in recent decades by relaxing the diagnostic criteria of some conditions and by expanding the range of problems conceptualized as falling within the psychiatric domain. Haslam documented how similar semantic inflation had occurred for concepts including abuse, addiction, prejudice, and trauma. Haslam proposed that these diverse concepts shared a focus on harm (i.e., the experience or infliction of actual or potential suffering). He therefore speculated that the correlated broadening of the creeping concepts reflected a rising sensitivity to harm within Western cultures.

In the present research we examine the following putatively creeping concepts:

① **Addiction.** This concept originally referred to physiological dependency on an ingested substance, but is increasingly used to identify psychological compulsion to engage in non-ingestive behaviors such as gambling or shopping.

② **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 power imbalance. More recent definitions extend bullying to adult workplace settings and relax the repetition,

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 29–34
Florence, Italy, August 2, 2019.

intentionality, and power imbalance criteria.

③ **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.

④ **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 non-conscious prejudice, and indicate that it may be manifest as anxiety or condescension rather than hostility.

⑤ **Trauma.** Four decades ago only personally encountered life-threatening events that are outside the realm of normal experience were recognized as traumatic by psychologists. More recent definitions include vicarious or indirect experiences of stressful events, including those that are relatively prevalent.

3 Related Work

Existing work on concept creep is primarily theoretical and the idea has been taken up by influential writers. Lukianoff and Haidt (2018) 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) has explored the attributes of people who hold relatively broad creeping-related concepts, finding that they tend to be politically liberal 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. However, to date no research has examined in theory's core claim that the meaning of harm-related concepts have systematically broadened within psychological discourse. The present research aims to remedy this lack using a large new corpus and diachronic language modelling.

Although diachronic studies of language have long history in linguistics, computational approaches to diachronic language modelling were introduced only recently. Jurgens and 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. (2016) then proposed the first prediction-based neural models. The latter work also formulated a number of laws of semantic change by exploring correlations between semantic changes and word frequency. Some of the laws were afterwards questioned and reformulated in (Dubossarsky et al., 2017).

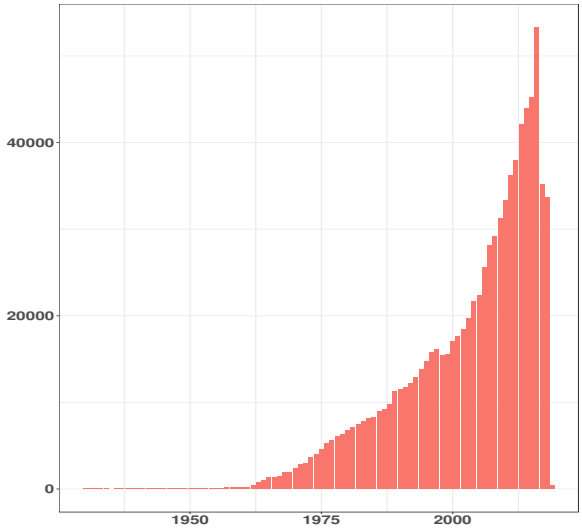

Figure 1: Statistics on the number of abstracts per year.

4 Corpus Description

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 in total. 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.[1] Fig. 1 presents the number of abstracts for each year . Due to 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.

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

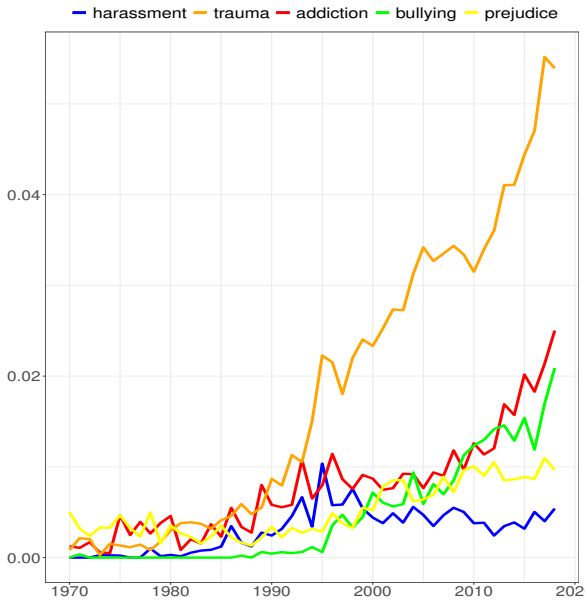

Figure 2: Relative word frequencies based on abstracts from psychology journals.

5 Experiments

5.1 Preprocessing Steps

We tokenized the corpus, removed punctuation, numbers, stop-words and non-English words, did fold-casing and lemmatization using SpaCy.[2]

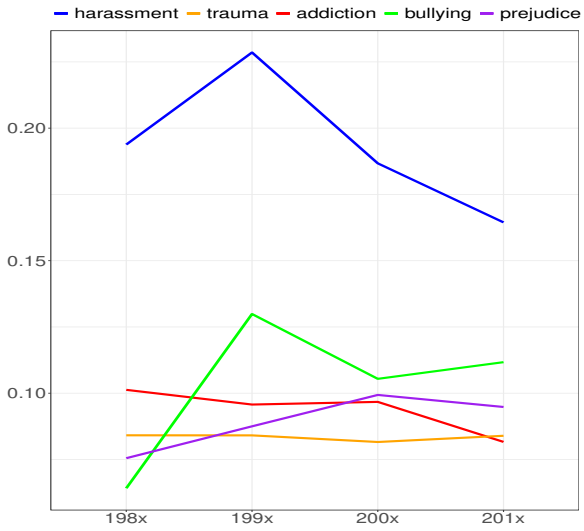

Figure 3: Average cosine similarities over five decades.

5.2 Frequency Analysis

For each of five concepts we first evaluated their (unigram) frequency distribution over time.[3] Although all the concepts demonstrate a certain relative raise of frequency, *Trauma* exhibits the steepest slope, while *Harassment* has its peak in the mid-nineties. Does it mean that *Trauma* became broader over time, i.e. it expanded to a whole range of new contexts? Has *Harassment* expanded to new contexts as well?

In the next section, we adapt two most widely used contemporary models, a count-based model from Sagi et al. (2009) and a prediction-based one from Hamilton et al. (2016). The former provides us with a time-specific measure of semantic breadth for each concept while the latter shows *how exactly* concepts changed. Both models have previously shown their utility at capturing semantic changes over time (Tahmasebi et al., 2018; Kutuzov et al., 2018).

5.3 Sagi et al.'s Model

Our first part of the experiments is based on the LSA-based model proposed by Sagi et al. (2009). We follow their instructions, i.e. we create a term–document co-occurrence matrix on the basis of the whole corpus. The total number of terms is restricted to 40,000 most frequent ones. We follow the vanilla TF-IDF model weights with logarithmic smoothing. The resulting matrix is factorized with SVD and truncated to 200 dimensions.[4] The resulting word embeddings are then contextualized for each decade starting 1980 and finishing 2019.[5] More specifically, in order to obtain a word vector representation for a certain decade, we randomly sample a number of its sentential occurrences[6] from that period, then extract contextual words at the pre-set window size.[7] The final sentence-specific representation is a bag-of-words, i.e. it is an average over corresponding context words representations. To estimate semantic breadth of a word, we evaluate pair-wise cosine similarities across all the sentence-specific representations. To reduce any biases, we repeat the above sampling process 10 times. Fig. 3 shows that concepts behave differently over time. For instance, *Trauma*, although being more frequently used, has not undergone significant changes in its meaning and stayed quite a "broad" concept. The notion of *Harassment*, on

[2] https://spacy.io/

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

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

[5] We only start with 1980s since certain concepts such as *bullying* were only introduced in 1970s, and the amount of data for them is insufficient for such an analysis.

[6] We set the number to 50

[7] We set the window size to 7

the other hand, was developing until 1990s where it reached the highest similarities in its contextual usages (became more semantically narrow). And during the last three decades the concept became broader again. Similarly, the concept of *Bullying* has been developed before 1990s, and then changed in both ways, becoming broader in 2000s and then narrowing down again in 2010s. Finally, during 2000s *Addiction* has expanded to new contexts such as "internet" and "smartphone". We will further study the changes in the next section.

	1980s-90s	1990s-00s	2000s-10s
addiction	0.35	0.23	0.23
bullying	0.64	0.27	0.19
harassment	0.29	0.21	0.18
prejudice	0.31	0.26	0.16
trauma	0.31	0.18	0.09

Figure 4: Cosine distances between decades.

5.4 Hamilton et al.'s Model

In order to investigate semantic change in a greater detail, we adapt a diachronic model from Hamilton et al. (2016). More specifically, we train a single word2vec model (Mikolov et al., 2013) for each time period, and then align them using the orthogonal Procrustes.[8] Following Hamilton et al. (2016), we consider two metrics to evaluate semantic changes over time:

1. Semantic displacement that shows to what extent an individual word has semantically changed during a certain time period. This is quantified as cosine distance between the aligned word embeddings from the corresponding time periods, i.e. cos-dist$(\mathbf{w}^t, \mathbf{w}^{t+\delta})$. Fig. 4 shows the results of evaluation and confirms our observations made earlier using the model from Sagi et al. (2009).

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

[8] Due to insufficient amount of data for earlier time periods, we train the models only on the following time frames: 1980-1989, 1990-1999, 2000-2010, 2011-2019.

	198x	199x	200x	201x
addiction.alcohol	0.31	0.31	0.3	0.21
addiction.cigarette	0.2	0.11	0.09	0.15
addiction.drug	0.4	0.3	0.34	0.28
addiction.gaming	0.04	0.18	0.18	0.39
addiction.heroin	0.47	0.35	0.33	0.21
addiction.internet	0	0.15	0.22	0.33
addiction.marijuana	0.37	0.26	0.18	0.17
addiction.narcotic	0.44	0.31	0.4	0.26
addiction.nicotine	0.15	0.28	0.23	0.21
addiction.opiate	0.39	0.27	0.38	0.29
addiction.sexual	0.01	0.17	0.15	0.12
addiction.smartphone	0	0	-0.07	0.21
bullying.child	0.13	0.2	0.2	0.13
bullying.peer	0.18	0.31	0.41	0.43
bullying.school	0.16	0.34	0.34	0.35
bullying.victim	0.18	0.34	0.49	0.46
bullying.workplace	0.09	0.26	0.39	0.4
harassment.cyber	0	-0.01	0.3	0.43
harassment.ethnic	0.07	0.18	0.16	0.17
harassment.gender	0.1	0.21	0.2	0.2
harassment.online	-0.13	0.1	0.18	0.25
harassment.peer	0.01	0.12	0.21	0.26
harassment.racial	0.18	0.25	0.32	0.31
harassment.sexual	0.18	0.16	0.15	0.25
harassment.student	0.12	0.18	0.19	0.18
harassment.woman	0.2	0.24	0.22	0.2
harassment.workplace	0.21	0.45	0.41	0.39
prejudice.black	0.42	0.34	0.35	0.33
prejudice.discrimination	0.14	0.3	0.32	0.44
prejudice.ethnic	0.41	0.44	0.38	0.4
prejudice.gay	0.28	0.31	0.27	0.36
prejudice.racial	0.48	0.5	0.52	0.53
prejudice.sex	0.24	0.22	0.18	0.12
prejudice.social	0.29	0.26	0.28	0.23
prejudice.woman	0.2	0.15	0.11	0.13
trauma.childhood	0.37	0.36	0.31	0.28
trauma.physical	0.19	0.15	0.09	0.03
trauma.psychological	0.19	0.25	0.31	0.28
trauma.sexual	0.11	0.2	0.24	0.19
trauma.stress	0.29	0.31	0.34	0.4

Figure 5: Cosine similarities over four decades.

the list for each decade. Fig. 5 presents a sample of nearest neighbors (words with highest cosine similarity) at a certain period of time and reflects changes of semantics of each concept. For instance, for *Addiction* it demonstrates a shift from substance-related concept in 1980s to behaviour-related one in 2010s. More specifically, we observe that it moved from "drug" and "narcotic"-related meaning towards "gaming", "internet", and "smartphone". *Bullying* has become more "victimized" and associated with workplace while its similarity to "school" and "child" stayed the same. Workplace also started being more related to *Harassment*, although, at the same time, its meaning expanded towards "cyber" and "online". Similarly, for *Trauma* we observe a shift from "physical" to "psychological" as well as an increase of a "stress" meaning. Finally, *Prejudice* has made strong connections to "discrimination" and "racial" while overall reduced for "black" and "woman".

6 Conclusion

The findings of our analyses illuminate and add nuance to our understanding of concept creep within academic psychology. The LSA-based analysis indicated that a sample of harm-related concepts have not undergone a consistent or linear pattern of semantic broadening. Since the 1990s *Addiction*, *Bullying* and *Harassment* have broadened, as the theory of concept creep would suggest, but the breadth of *Trauma* has been relatively static and *Prejudice* has somewhat narrowed. The analysis of semantic displacement points to a more consistent diachronic pattern: all five concepts changed most substantially from the 1980s to the 1990s and changed progressively less thereafter. This finding implies that the final two decades of the 20th century are 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., cyberharassment), some have added new semantic domains (e.g., *Addiction* incorporating non-ingestive behaviors such as gaming and smartphone use), and others have shifted emphasis (e.g., *Trauma* becoming associated less with physical injury and more with psychological stress).

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 do not adequately capture how meanings may shift "vertically" to encompass less severe phenomena.

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. 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.

References

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.

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

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

David Jurgens and Keith Stevens. 2009. Event detection in blogs using temporal random indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 9–16. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *ACL 2014*, page 61.

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.

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

Greg Lukianoff and Jonathan Haidt. 2018. *The Coddling of the American Mind: How Good Intentions and Bad Ideas Are Setting Up a Generation for Failure*. Penguin UK.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Eyal Sagi, Stefan Kaufmann, and 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*, pages 104–111. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

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.