Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang

Anonymous ACL submission

Abstract

All living languages are continually undergoing changes, and the mechanisms that underlie language change are still a matter of debate. In this work, we approach language change through the lens of causality in order to model not only how various distributional factors associate with language change, but how they causally affect it. In particular, we study slang, which is an informal language that is typically restricted to a specific group or social setting. We analyze the semantic change and frequency shift of slang words and compare them to those of standard, nonslang words. With causal discovery and causal inference techniques, we measure the effect that word type (slang/nonslang) has on both semantic change and frequency shift, as well as its relationship to frequency, polysemy and part of speech. Our analysis provides some new insights in the study of semantic change, e.g., we show that slang words undergo less semantic change but tend to have larger frequency shifts over time.[1]

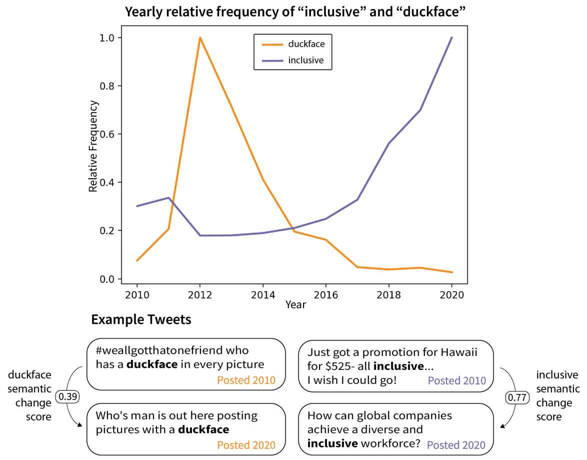


Figure 1: We observe very different change dynamics for the slang word "duckface" and the nonslang word "inclusive". "Inclusive" has acquired a new meaning, reflected in a high semantic change score of 0.77 as measured by our model. "Duckface" undergoes little semantic change, scored 0.39 by our model, while its usage frequency varies greatly.

1 Introduction

Language is a continuously evolving system, constantly resculptured by its speakers. The forces that drive this evolution are many, ranging from phonetic convenience to sociocultural changes (Blank, 1999). In particular, the meanings of words and the frequencies in which they are used are not static, but rather evolve over time.

Several previous works, in both historical and computational linguistics, have described diachronic mechanisms, often suggesting causal relationships. For example, semantic change, i.e. change in the meaning of a word, has both been suggested to *cause* (Wilkins, 1993; Hopper and Traugott, 2003) and *be caused by* (Hamilton et al., 2016) polysemy, while also part of speech (POS) has been implied to be a causal factor behind semantic change (Dubossarsky et al., 2016). However, none of these studies perform a causal analysis to verify these claims of causal relationships. Causality allows us to not only infer causal effects between pairs of variables, but also model their interactions with other related factors.

In this work, we focus on the linguistic evolution of slang, defined as colloquial and informal language commonly associated with particular groups (González, 1998; Bembe and Beukes, 2007), and use a causal framework to compare the change dynamics of slang words to those of standard language. More specifically, we compare the *semantic change* as well as the changes in frequency, i.e. *frequency shift*, over time between slang words and standard, nonslang words. We learn a *causal graphical model* (Spirtes et al., 2000) to assess how these variables interact with other factors they have

[1]Code will be published with the camera-ready version.

been previously found to correlate with, such as *frequency*, *polysemy* and *part of speech*. Having discovered a graph, we proceed to use *do-calculus* (Pearl, 1995) to evaluate the causal effects of a word's *type* (slang/nonslang) on semantic change and frequency shift.

Semantic change is measured using the average pairwise distance (APD) (Sagi et al., 2009; Giulianelli et al., 2020) between time-separated contextualized representations, which were obtained from a Twitter corpus via a bi-directional language model (Liu et al., 2019). Our metric builds on recent semantic change literature (Schlechtweg et al., 2020), with novel additions of dimensionality reduction and a combined distance function.

By deploying a causal analysis, we establish that there is not just an association, but a direct effect of a word's type on its semantic change and frequency shift. We find that a word being *slang* causes it to undergo slower semantic change and more rapid decreases in frequency. To illustrate, consider the slang word "duckface" and the nonslang word "inclusive" as shown in Figure 1. Our analysis also sheds light on a couple of previous findings in the diachronic linguistics literature. We find support for the S-curve theory (Kroch, 1989), showing a causal effect from a word's polysemy to its frequency. This relationship is evident in the increase in frequency that the word "inclusive" displays in Figure 1 after it develops a new meaning (Merriam-Webster, 2019). However, similar to Dubossarsky et al. (2017), we do not find a causal link to semantic change from frequency, polysemy or POS as suggested in previous works (Hamilton et al., 2016; Dubossarsky et al., 2016).

In summary, our main contributions are threefold: (i) we introduce tools from the causality literature in order to analyze change dynamics in language; (ii) we propose a semantic change metric using contextualized word representations and (iii) we discover some interesting insights about slang words and semantic change – e.g. showing that the change dynamics of slang words are different from those of nonslang words, exhibiting both more rapid frequency fluctuations and less semantic change.

2 Related Work

2.1 Semantic Change

A typical method for measuring semantic change is by comparing word representations across time periods (Gulordava and Baroni, 2011; Kim et al., 2014; Jatowt and Duh, 2014; Kulkarni et al., 2015; Eger and Mehler, 2016; Schlechtweg et al., 2019). With this approach, previous research has proposed laws relating semantic change to other linguistic properties. For instance, Dubossarsky et al. (2016) find that verbs change faster than nouns, whereas Hamilton et al. (2016) discover that polysemous words change at a faster rate, while frequent words change slower. However, the validity of some of these results has been questioned via methods of case-control matching (Dubossarsky et al., 2017), highlighting the influence of word frequency when modeling change (Hellrich and Hahn, 2016). Such analyses can indeed help give stronger evidence for causal effects. In this work we take a methodologically different approach, considering observational data alone for our causal analysis.

The aforementioned approaches rely on fixed word representations. Limited by assigning one vector to each word, fixed embeddings may fail to capture polysemous words properly, as well as certain contextual nuances. More recent approaches (Hu et al., 2019; Giulianelli et al., 2020) have highlighted the limitations of using fixed representations and proposed semantic change measures based on contextualized word embeddings (Peters et al., 2018; Devlin et al., 2019). This has lead to a further stream of work on semantic change detection with contextualized embeddings (Martinc et al., 2020; Kutuzov and Giulianelli, 2020; Schlechtweg et al., 2020; Montariol et al., 2021; Kutuzov et al., 2021; Laicher et al., 2021). We build upon this line of work and extend them using PCA and a combination of distance metrics.

2.2 Characterization and Properties of Slang

Slang is an informal, unconventional part of the language, often used in connection to a certain setting or societal trend (Dumas and Lighter, 1978). It can reflect and establish a sense of belonging to a group, (González, 1998; Bembe and Beukes, 2007; Carter, 2011) or to a generation (Citera et al., 2020; Earl, 1972; Barbieri, 2008).

Mattiello (2005) highlights the role slang plays in enriching the language with neologisms, and claims that it follows unique word formation processes. Inspired by this, Kulkarni and Wang (2018) propose a data-driven model for emulating the generation process of slang words that Mattiello (2005) describes. Others have described the ephemerality of slang words (González, 1998; Carter, 2011), although this property has not been previously verified by computational approaches.

3 Causal Methodology for Change Dynamics

Examining change dynamics through a causal lens helps to determine the existence of direct causal effects, by modeling the interactions between variables. In this section, we first give a short overview on relevant work on causality, before presenting how we apply these concepts to word change dynamics.

3.1 Overview of Causal Discovery and Causal Inference

A common framework for causal reasoning is through *causal directed acyclic graphs* (DAGs) (Pearl, 2009). A causal DAG consists of a pair (G, P) where $G = (V, E)$ is a DAG and P is a probability distribution over a set of variables. Each variable is represented by a node $v \in V$, and the graph's edges $e \in E$ reflect causal relationships. There are two main tasks in causality. *Causal discovery* is the task of uncovering the causal DAG that explains observed data. Assuming a causal DAG, the task of *causal inference* then concerns determining the effect that intervening on a variable, often referred to as *treatment*, will have on another variable, often referred to as *outcome*.

The causal DAG is often inferred from domain knowledge or intuition. However, in cases where we cannot safely assume a known causal structure, causal discovery methods come in useful. Constraint-based methods (Spirtes et al., 2000) form one of the main categories of causal discovery techniques. These methods use conditional independence tests between variables in order to uncover the causal structure. To do so, they rely on two main assumptions: the global Markov assumption and the faithfulness assumption. Together they state that we observe conditional independence relations between two variables in the distribution if and only if these two variables are d-separated (Geiger et al., 1990) in the graphical model. For more details, we refer to Appendix D.1.

Causal inference is commonly approached with do-calculus (Pearl, 1995). We denote the intervention distribution $\mathbb{P}(Y|do(X = x))$ to be the distribution of the outcome Y conditioned on an intervention $do(X = x)$ which forces the treatment variable X to take on the value x. Note that this is in general not necessarily equal to $\mathbb{P}(Y|X = x)$. When they are not equal, we say that there is *confounding*. Confounding occurs when there is a third variable Z, which causes both the treatment X and the outcome Y.

We say that there is a causal effect of X on Y if there exist x and x' such that

$$\mathbb{P}(Y|do(X = x)) \neq \mathbb{P}(Y|do(X = x')) \,. \qquad (1)$$

One way to quantify the causal effect is with the *average causal effect (ACE)*:

$$\mathbb{E}[Y|do(X = x)] - \mathbb{E}[Y|do(X = x')] \,. \qquad (2)$$

To estimate the causal effect using observational data, we need to rewrite the intervention distribution using only conditional distributions. Assuming a causal DAG, this can be done with the *truncated factorization formula* (Pearl, 2009),

$$\begin{aligned} &\mathbb{P}(X_V|do(X_W = x_W)) = \\ &= \prod_{i \in V \setminus W} \mathbb{P}(X_i|X_{pa(i)}) \mathbb{1}_{\{X_W = x_W\}} \,, \end{aligned} \qquad (3)$$

for $W \subset V$.

3.2 Causality for Change Dynamics

In this work, we estimate the direct causal effect of a word's type on its semantic change and frequency shift dynamics. In order to establish that such an effect exists, and to know which variables to control for, we turn to causal discovery algorithms. The variables in our causal graph additionally include frequency, polysemy and POS.

For learning the causal graph, we choose the constraint-based PC-stable algorithm (Colombo and Maathuis, 2014), an order-independent variant of the well-known PC algorithm (Spirtes et al., 2000), discussed in Appendix D.1. We are learning a mixed graphical model (Lauritzen, 1996; Lee and Hastie, 2015), consisting of both continuous (e.g. frequency) and categorical (e.g. type) variables.

Having learned the causal graph (Section 6.2), we proceed to estimate the ACE of word type on both semantic change and frequency shift using do-calculus (Section 6.3).

4 Slang and Nonslang Word Selection

We select 100 slang words and 100 nonslang words for our study, presented in Appendix E. In the tradeoff between statistical significance and time spent on computation and data collection, we found that a set of 200 words was enough to get highly significant results. However, we note that our methodology is general and can be applied to a larger set of words. The slang words are randomly sampled from the Online Slang Dictionary,[2] which provides well-maintained and curated slang word definitions as well as a list of 4,828 featured slang words as of June 2021. The scope of our study encompasses single-word expressions, and as such we filter out 2,169 multi-word expressions. To further clean the data, we also delete words with only one character and acronyms. Lastly, we limit the causal analysis to words that are exclusively either slang or nonslang, excluding "hybrid" words with both slang and nonslang meanings, such as "kosher" or "tool". Including words of this type would have interfered with the causal analysis by creating a hardcoded dependency between word type and polysemy, as these words by definition are polysemous. We do however perform a separate analysis of the hybrid words in Appendix C.

For the reference set of standard, nonslang, words we sample 100 words uniformly at random from a list of all English words, supplied by the wordfreq library in Python (Speer et al., 2018).

5 Data Collection

We curate a Twitter dataset from the years 2010 and 2020, which we select as our periods of reference, and collect the following variables:

- **Word type:** Whether a word is slang or not
- **Word frequency:** The average number of tweets containing the word per day in 2010 and 2020 (Section 5.2)
- **Frequency Shift:** The relative difference in frequency the word has undergone between 2010 and 2020 (Section 5.3)
- **Polysemy:** The number of senses a word has (Section 5.4)
- **Part of speech:** A binary variable for each POS tag (Section 5.5)
- **Semantic change:** The semantic change score of the word from 2010 to 2020 (Section 5.6)

5.1 Twitter Dataset

As a social media platform, Twitter data is rich in both slang and nonslang words. The Twitter dataset we curated comprises 170,135 tweets from 2010 and 2020 that contain our selected words. Sampling tweets from two separate time periods allows us to examine the semantic change over a 10-year gap. For every slang and nonslang word, and each of the two time periods, we obtain 200-500 random tweets that contain the word and were posted during the corresponding year. We keep each tweet's text, tweet ID, and date it was posted. As a post-processing step, we remove all URLs and hashtags from the tweets. To protect user privacy, we further replace all user name handles with the word "user." On average, we have 346 tweets per slang word and 293 tweets per nonslang word.

[2] http://onlineslangdictionary.com/

5.2 Word Frequency

We approximate a word's frequency by the average number of times it is tweeted within 24 hours. This average is calculated in practice over 40 randomly sampled 24 hour time frames in a given year, in each of which we retrieve the number of tweets containing the word. The frequencies are calculated separately for 2010 and 2020. Due to the growing popularity of social media, the number of tweets has significantly increased over the decade. Therefore, we divide the counts from 2020 by a factor of 6.4, which is the ratio between the average word counts in both years in our dataset. The frequencies from both years are then averaged to provide the *frequency* variable for the causal analysis.

5.3 Frequency Shift

We are now interested in analyzing the dynamics of frequency shifts. To evaluate the relative change in frequency for a given word w we take

$$\text{FreqShift}(w) = \log \frac{x_{2020}(w)}{x_{2010}(w)} \tag{4}$$

where, $x_k(w)$ is the frequency of word w in year k. This has been shown to be the only metric for relative change that is symmetric, additive, and normed (Tornqvist et al., 1985). Importantly, this measure symmetrically reflects both increases and decreases in relative frequency. The mean relative changes in frequency were $-0.486(\pm 1.644)$ for slang words and $0.533(\pm 1.070)$ for nonslang words, where a positive score corresponds to an increase in frequency. As evident in Figure 2, not only did more slang words exhibit a decrease in frequency than nonslang ones, the words that showed the highest frequency increase are also slang.

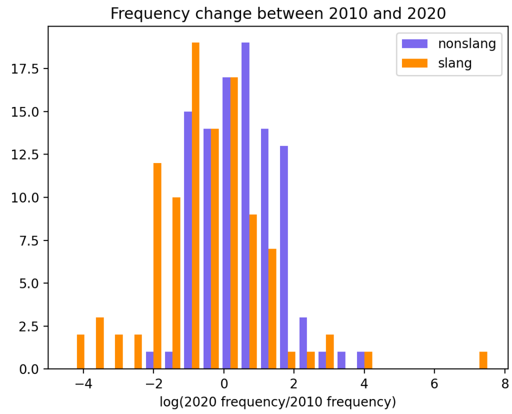


Figure 2: Relative shift in frequency from 2010 to 2020, where a positive score corresponds to an increase in frequency. We see that slang words present both the highest increases and the highest decreases in frequency. Moreover, a large frequency decrease is observed exclusively in a set of slang words, indicating these words faded from usage during the decade.

We also examine the absolute value of Eq. (4) to evaluate the degree of change, may it be a decrease or an increase. We find that, as expected, slang words have significantly higher changes in frequency than nonslang words ($p < 0.05$). See Appendix C for more details.

5.4 Polysemy

We define a word's polysemy score as the number of distinct senses it has[3]. For nonslang words we take the number of senses the word has in Merriam Webster and for slang words we take the number of definitions on the Online Slang Dictionary. We use two separate resources as we find that no dictionary encapsulates both slang and nonslang words. The mean polysemy scores are (2.074 ± 2.595) for slang words and (3.079 ± 2.780) for nonslang words with a significant difference in distribution ($p < 0.05$) according to a permutation test, implying that the latter are used with a larger variety of meanings. In addition, the slang senses of the hybrid words exhibit a distribution similar to those of the slang words (Appendix C). More polysemous words tend to have a higher word frequency in our dataset – the log transform of frequency and polysemy display a highly significant ($p < 0.001$) linear correlation coefficient of 0.350.

[3]Note that this definition also encapsulates potential cases of homonymy. We choose not to make a distinction between polysemy and homonymy in this analysis.

5.5 Part of speech

For each word, we retrieve four binary variables, indicating whether a word can be used as noun, verb, adverb or adjective, which were the four major POS tags observed in our data. To calculate these variables we run the NLTK POS tagger (Loper and Bird, 2002) on the tweets, and collect the distribution of POS tags for each word. Note that a word may have more than one POS tag, depending on the context in which it is used. Each of the binary variables is then set to be 1 if the word had the corresponding POS tag in at least 5% of its tweets and 0 otherwise.

5.6 Semantic Change Score

In this section we explain the details of how we obtain the semantic change scores. We start by fine-tuning a bi-directional language model on a slang-dense corpus (Section 5.6.1), after which we survey the literature and propose metrics (Section 5.6.2) that we use to perform an extensive experimentation study to find the most suitable one (Section 5.6.3). Finally, we apply this metric to our sets of slang and nonslang words on the Twitter data (Section 5.6.4).

5.6.1 Obtaining Contextualized Representations

We familiarize the bi-directional language model with slang words and the contexts in which they are used by fine-tuning it on the masked language modeling task. For this purpose we use a web-scraped dataset from the Urban Dictionary, previously collected by Wilson et al. (2020). After preprocessing and subsampling, the details of which can be found in Appendix A.1, we are left with a training set of $200,000$ slang-dense text sequences.

As our bi-directional language model we select RoBERTa (Liu et al., 2019). Beyond performance gains compared to the original BERT (Devlin et al., 2019), we select this model since it allows for more subword units. We reason that this could be useful in the context of slang words since potentially some of the sub-units used in these words would not have been recognized by BERT. We choose the smaller 125M parameter base version for computational reasons. We train the model using the Adam optimizer (Kingma and Ba, 2015) with different learning rates γ. The lowest loss on the test set was found with $\gamma = 10^{-6}$, which we proceed with for scoring semantic change. For more details on training configurations, we refer to Appendix A.2.

5.6.2 Quantifying Semantic Change

In order to select a change detection metric, we evaluate our model on the SemEval-2020 Task 1 on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020). This task provides the first standard evaluation framework for semantic change detection, using a large-scale labeled dataset for four different languages. We restrict ourselves to English and focus on subtask 2, which concerns ranking a set of 37 target words according to their semantic change between two time periods. The ranking is evaluated using Spearman's rank-order correlation coefficient ρ.[4] Our space of configurations includes layer representations, dimensionality reduction techniques and semantic change metrics.

Layer Representations: Previous work (Ethayarajh, 2019) has shown that embeddings retrieved from bi-directional language models are not isotropic, but are rather concentrated around a high-dimensional cone. Moreover, the level of isotropy may vary according to the layer from which the representations are retrieved (Ethayarajh, 2019; Cai et al., 2021). This leads us to experiment with representations from different layers in our fine-tuned RoBERTa model, namely, taking only the first layer, only the last layer or summing all layers.

Dimensionality Reduction: To the best of our knowledge, only one previous semantic change detection approach (Rother et al., 2020) has incorporated dimensionality reduction, more specifically UMAP (McInnes et al., 2018). In addition to UMAP, we also experiment with PCA.

Metrics for Semantic Change: Given representations $\mathcal{X}_t = \{\boldsymbol{x}_{1,t}, ..., \boldsymbol{x}_{n_t,t}\}$ for a particular word in time period t, we define the average pairwise distance (APD) between two periods as

$$\mathrm{APD}(\mathcal{X}_{t_1}, \mathcal{X}_{t_2}) = \frac{1}{n_{t_1} n_{t_2}} \sum_{\substack{\boldsymbol{x}_{i,t_1} \in \mathcal{X}_{t_1} \\ \boldsymbol{x}_{j,t_2} \in \mathcal{X}_{t_2}}} d(\boldsymbol{x}_{i,t_1}, \boldsymbol{x}_{j,t_2}) \, , \tag{5}$$

for some distance metric $d(\cdot,\cdot)$, where n_{t_1}, n_{t_2} are the number of words in each time period. We experiment with Euclidean distance $d_2(\boldsymbol{x}_1, \boldsymbol{x}_2)$, cosine distance $d_{\cos}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and Manhattan distance $d_1(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Furthermore, we propose a novel combined metric. Note that $d_2(\cdot,\cdot) \in [0, \infty]$ and $d_{\cos}(\cdot,\cdot) \in [0, 2]$. Further note that

$$||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2^2 \leq ||\boldsymbol{x}_1||_2^2 + ||\boldsymbol{x}_2||_2^2 \tag{6}$$

Normalizing both metrics for a support in $[0, 1]$, we get a combined metric with the same unit support to be the following average:

$$d_{2,\cos}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{0.5 \cdot d_2(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sqrt{||\boldsymbol{x}_1||^2 + ||\boldsymbol{x}_2||^2}} + \frac{d_{\cos}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{4} \tag{7}$$

We argue that this provides a more complete metric, capturing both absolute distance and the angle between vectors.

In addition to the APD metrics, we experiment with distribution-based metrics (see Appendix B.1).

[4] We note the caveat that our model is fine-tuned on Urban Dictionary text, while the older of the two English datasets of SemEval consists of text from 1810-1860.

Reduction	h	APD	Score
PCA	100	d_2 and $d_{\cos}$	**0.489****
PCA	100	$d_{\cos}$	0.464**
PCA	100	d_2	0.298
None	768	d_2 and $d_{\cos}$	0.345*

Table 1: Spearman's rank-order correlation coefficients between our semantic change scores and the ground truth across different dimensionality reduction techniques for APD (*: $p < 0.05$, **: $p < 0.01$).

5.6.3 Evaluating the Semantic Change Scores

We first compare the results for the three types of layer representations different APD metrics, and note that summing all layer representations yields the best results. Consequentially, we proceed with the rest of the experiments using only these representations. For both PCA and UMAP, we experiment with projecting the representations down to $h \in \{2, 5, 10, 20, 50, 100\}$ dimensions. These combinations are tested together with the APD metrics as presented in Section 5.6.2 as well as the distribution-based metrics described in Appendix B. The latter do not however in general display significant ($p < 0.05$) correlations.

We present a small subset of the scores resulting from the APD configurations in Table 1, highlighting our finding that both PCA dimensionality reduction and using a combined the metrics and improve the performance. More results and comparisons to baselines are presented in Appendix B.3. We observe that the proposed combined metric consistently outperforms both d_2 and d_{cos} across values

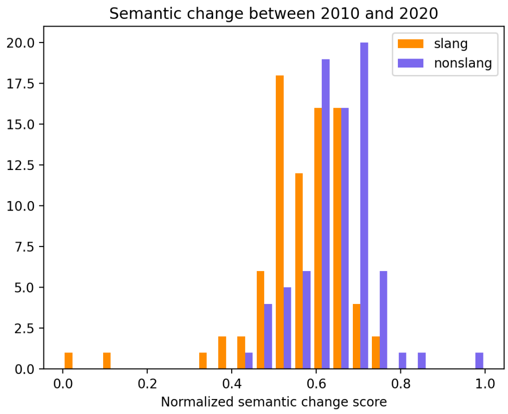


Figure 3: Semantic change scores between 2010 and 2020. We see that nonslang words typically underwent larger changes in meaning throughout the decade.

of h for PCA. We also note that UMAP projections perform poorly with the APD metrics and that projecting down to 50-100 dimensions seems to be optimal, which maintains 70-85% of the variance as we illustrate in Appendix B.2. In addition, both norm-based metrics d_1 and d_2 perform worse with dimensionality reduction. As our final metric, we choose the best performing configuration on SemEval, with PCA $h = 100$ and the combined metric, as seen in Table 1.

5.6.4 Semantic Change Scores for Slang and Nonslang Words on the Twitter Dataset

We obtain semantic change scores using the Twitter dataset described in Section 5.1. For the semantic change analysis, we exclude words that have less than 150 tweets in each time period within the dataset, which leaves us with 80 slang and 81 nonslang words. We also normalize the scores according to the sample. The resulting semantic change scores are shown in Figure 3. The mean semantic change scores are $0.564(\pm 0.114)$ for slang words and $0.648(\pm 0.084)$ for nonslang words. The difference in semantic change score distributions is significant ($p < 0.001$) via a permutation test. The word with the highest semantic change score of 1 is “anticlockwise”, and the word with the lowest score of 0 is “whadja”.

6 Causal Analysis

6.1 Preparation for Causal Discovery

PC-stable is constraint-based and thus makes use of conditional independence tests. In the case of continuous Gaussian variables, we can perform partial correlation tests to assess conditional independence, since zero partial correlation in this case is equivalent to conditional independence (Baba et al., 2004). As word frequency has been suggested to follow a lognormal distribution (Baayen, 1992), we take the log transform of it. The continuous variables *semantic change*, *frequency change* and *log-frequency* are then all assumed to be approximated well by a Gaussian distribution, which is confirmed by diagnostic density and Q-Q plots (displayed in Appendix D.2).

We categorize the discrete polysemy variable, experimenting with nine different plausible categorizations for the sake of robustness of the results. Word type and POS are categorical in nature. For the categorical variables and for mixes of categorical and continuous variables, we perform chi-squared mutual information based tests (Edwards, 2000), since the approximate null distribution of the mutual information is chi-squared (Brillinger, 2004). For all conditional independence tests we experiment with significance levels $\alpha \in \{0.01, 0.03, 0.05\}$.

6.2 Resulting Causal Structure

In Figure 4 we see the result from the above approach, with dotted lines representing edges that were apparent in most but not all of the configurations. See Appendix D.3 for a sensitivity analysis.

We first observe that word type has a direct causal effect on both the semantic change score and the frequency shift, without any confounders. We also note that none of the four POS categories, which are all gathered in one node in Figure 4, have a causal link to any of the other variables. We additionally observe a dependency between word type and polysemy. This edge could not be oriented by the PC-stable algorithm, however we manually orient it as outgoing from type and ingoing to polysemy, since an intervention on type should have a causal effect on the number of word senses and not vice versa. It is also interesting to note that polysemy does not seem to have a causal effect on semantic change. Its association with semantic change ($p < 0.05$, rejecting the null hypothesis of independence between polysemy and semantic change) is instead confounded by word type.

6.3 Causal Effects

In our case of no confounders, evaluating the ACE of word type on semantic change is straightforward, as it reduces to the difference between the

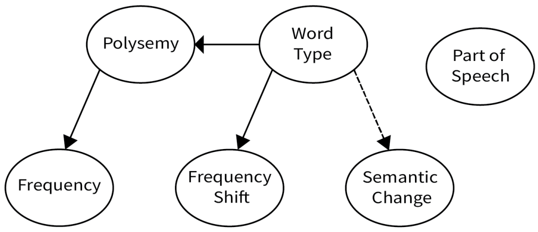


Figure 4: DAG representing the causal relationships in our dataset.

conditional expectations:

$$\mathbb{E}[S|do(T=\text{nonslang})] - \mathbb{E}[S|do(T=\text{slang})] = \\ = \mathbb{E}[S|T=\text{nonslang}] - \mathbb{E}[S|T=\text{slang}] \tag{8}$$

See Appendix D.4 for a derivation. The case of frequency shift is analogous.

We estimate the expectations by the sample means on the normalized values and get an average causal effect of 0.084, which is a highly significant value ($p < 0.001$) based on a t-test. For the observed changes in relative frequency, calculated according to Eq. (4), we get an average causal effect of 1.017 ($p < 0.001$ via a t-test).

7 Discussion

We analyze the dynamics of frequency shift and semantic change in slang words, and compare them to those of nonslang words. Our analysis shows that **slang words change slower in semantic meaning, but adhere to more rapid frequency fluctuations, and are more likely to greatly decrease in frequency**. Our study is the first computational approach to confirm this property in slang words (González, 1998; Carter, 2011).

To ensure that this is the result of a causal effect, and not mediated through another variable or subject to confounders, we model the data with a causal DAG, by also considering the potential interacting variables polysemy, frequency and POS. We discover that there is no influence of confounders, nor are there mediators between a word's type and its semantic change or its frequency shift, which **confirms a direct causal effect**.

Our results are consistent with those of Dubossarsky et al. (2017), which found that neither the law relating semantic change to frequency, polysemy (Hamilton et al., 2016) nor prototypicality (Dubossarsky et al., 2015) were found to be as strong as previously thought after a case-control study using a scenario without semantic change. Indeed, there is no directed path from polysemy or frequency to semantic change in our causal graph, but they are both influenced by word type. We leave for future research to explore whether other categorizations of words sustain this result.

In addition, our analysis does not support the claim that POS could underlie semantic change (Dubossarsky et al., 2016). We note however that as our vocabulary contains 50% slang words, the results need not be consistent with results obtained with a word sample drawn from standard language.

Moreover, in the causal structure we discover that **word polysemy has a direct effect on word frequency**, which is in line with previous linguistic studies showing that a word's frequency grows in an S-shaped curve when it acquires new meanings (Kroch, 1989; Feltgen et al., 2017), as well as a known positive correlation between polysemy and frequency (Lee, 1990; Casas et al., 2019). We emphasize that this relationship is not merely an artifact of contextualized word representations being affected by frequency (Zhou et al., 2021), since our polysemy score does not rely on word representations as in Hamilton et al. (2016). Our approach is however not without drawbacks – the polysemy variable is collected from dictionaries, which may be subjective in their assignments of word senses.

Limitations: Our study, along with previous work on the dynamics of semantic change, are all limited by only considering distributional factors. For instance, linguists have suggested that sociocultural, psychological and political factors (Blank, 1999; Bochkarev et al., 2014) all drive word change dynamics, and slang words are not an exception. Returning to our example "duckface", it may be that its rapid decrease in frequency is also due to social factors (Miller, 2011). Phrased differently, our causal analysis is not immune to issues with missing variables. Nonetheless, we do believe that such a causal analysis provides a useful tool to understand the underlying mechanisms of language.

8 Conclusion

In this work, we have analyzed the diachronic mechanisms of slang language with a causal methodology. This allowed us to establish that a word's type has a direct effect on its semantic change and frequency shift, without mediating effects from other distributional factors.

Ethical Considerations

Our dataset is comprised solely of English text, and our analysis therefore applies uniquely to the English language, and results may differ in other languages. Moreover, for the purpose of this study, we curated a dataset of $170,135$ tweets. To protect the anonymity of users, we remove author IDs from the data, and replace all usernames with the general token "user." In the Urban Dictionary dataset we received from Wilson et al. (2020), we similarly remove the author IDs and only consider the entry text.

References

R. Harald Baayen. 1992. Statistical models for word frequency distributions: A linguistic evaluation. Computers and the Humanities, 26:347–363.

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. 2004. Partial correlation and conditional correlation as measures of conditional independence. Australian & New Zealand Journal of Statistics, 46(4):657–664.

Federica Barbieri. 2008. Patterns of age-based linguistic variation in american english1. Journal of Sociolinguistics, 12:58 – 88.

Magdeline Princess Bembe and Anne-Marie Beukes. 2007. The use of slang by black youth in gauteng. Southern African Linguistics and Applied Language Studies, 25(4):463–472.

Andreas Blank. 1999. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. In Historical Semantics and Cognition, pages 61–90, Berlin/New York. Mouton de Gruyter.

Vladimir Bochkarev, Valery Solovyev, and Søren Wichmann. 2014. Universals versus historical contingencies in lexical evolution. Journal of The Royal Society Interface, 11:1–23.

David R. Brillinger. 2004. Some data analyses using mutual information. Brazilian Journal of Probability and Statistics, 18(2):163–182.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In International Conference on Learning Representations.

Phillip M. Carter. 2011. Michael adams, slang: The people's poetry. oxford: Oxford university press. pp. 238. hb. 23.95. Language in Society, 40(3):400–401.

Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer i Cancho, and Jaume Baixeries. 2019. Polysemy and brevity versus frequency in language. Computer Speech Language, 58:19–50.

Maryalice Citera, Coreyann Spence, and Madalena Spero. 2020. Differences in emotional word use across generations in the united states. Journal of Business and Social Science Review, Vol. 1; No. 2.

Diego Colombo and Marloes H. Maathuis. 2014. Order-independent constraint-based causal structure learning. Journal of Machine Learning Research, 15(116):3921–3962.

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 and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In NetWordS.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2016. Verbs change more than nouns: a bottom-up computational approach to semantic change. Lingue e linguaggio, Rivista semestrale, (1/2016):7–28.

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.

Bethany K. Dumas and Jonathan Lighter. 1978. Is slang a word for linguists. American Speech, 53:5.

Kim Earl. 1972. Semantic influence and concept attainment of slang and its effects on parents' and teenagers' linguistic interaction. All Graduate Theses and Dissertations.

David Edwards. 2000. Introduction to Graphical Modelling, 2nd edition. Springer.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 52–58, Berlin, Germany. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Q. Feltgen, B. Fagard, and J.-P. Nadal. 2017. Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. Royal Society Open Science, 4(11):170830.

Dan Geiger, Thomas Verma, and Judea Pearl. 1990. Identifying independence in bayesian networks. Networks, 20(5):507–534.

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

Félix Rodríguez González. 1998. Reviews : Slang and sociability: In-group language among college students. by connie eble. chapel hill: University of north carolina press, 1996. xi + 228. Journal of English Linguistics, 26(3):247–265.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

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), pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2785–2796, Osaka, Japan. The COLING 2016 Organizing Committee.

Paul J. Hopper and Elizabeth Closs Traugott. 2003. Grammaticalization. Cambridge University Press, Cambridge, UK.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In IEEE/ACM Joint Conference on Digital Libraries, pages 229–238.

Jens Kaiser, Dominik Schlechtweg, Sean Papay, and Sabine Schulte im Walde. 2020. IMS at SemEval-2020 task 1: How low can you go? dimensionality in lexical semantic change detection. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 81–89, Barcelona (online). International Committee for Computational Linguistics.

Markus Kalisch and Peter Bühlmann. 2007. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. J. Mach. Learn. Res., 8:613–636.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and 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, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Anthony S. Kroch. 1989. Reflexes of grammar in patterns of language change. Language Variation and Change, 1(3):199–244.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In Proceedings of the 24th International Conference on World Wide Web, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Vivek Kulkarni and William Yang Wang. 2018. Simple models for word formation in slang. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1424–1434, New Orleans, Louisiana. Association for Computational Linguistics.

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

Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. 2021. Grammatical profiling for semantic change detection. In Proceedings of the 25th Conference on Computational Natural Language Learning, pages 423–434, Online. Association for Computational Linguistics.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 192–202, Online. Association for Computational Linguistics.

Steffen L. Lauritzen. 1996. Graphical models. Number 17 in Oxford Statistical Science Series. Clarendon Press.

Christopher J. Lee. 1990. Some hypotheses concerning the evolution of polysemous words. Journal of Psycholinguistic Research, 19:211–219.

Jason D. Lee and Trevor J. Hastie. 2015. Learning the structure of mixed graphical models. Journal of Computational and Graphical Statistics, 24(1):230–253. PMID: 26085782.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv, abs/1907.11692.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing evolution in word usage: Just add more clusters? In Companion Proceedings of the Web Conference 2020, WWW '20, page 343–349, New York, NY, USA. Association for Computing Machinery.

Elisa Mattiello. 2005. The pervasiveness of slang in standard and non-standard english. page 7–41. Mots Palabras Words 6.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Merriam-Webster. 2019. We added new words to the dictionary in september 2019.

Sarah Miller. 2011. Duck hunting on the internet. The New York Times.

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, Online. Association for Computational Linguistics.

Judea Pearl. 1995. Causal diagrams for empirical research. Biometrika, 82(4):669–688.

Judea Pearl. 2009. Causal inference in statistics: An overview. Statistics Surveys, 3(none):96 – 146.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. Elements of causal inference: foundations and learning algorithms. The MIT Press.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

David Rother, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 187–193, Barcelona (online). International Committee for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65.

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, Athens, Greece. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Gideon Schwarz. 1978. Estimating the Dimension of a Model. The Annals of Statistics, 6(2):461 – 464.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. Causation, prediction, and search. MIT press.

Leo Tornqvist, Pentti Vartia, and Yrjo O. Vartia. 1985. How should relative changes be measured? The American Statistician, 39(1):43–46.

David P. Wilkins. 1993. From part to person: Natural tendencies of semantic change and the search for cognates. Cognitive Anthropology Research Group at the Max Planck Institute for Psycholinguistics.

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang NLP applications. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4764–4773, Marseille, France. European Language Resources Association.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2021. Frequency-based distortions in contextualized word embeddings. CoRR, abs/2104.08465.

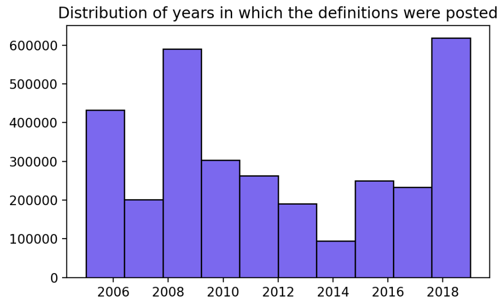


Figure 5: Frequency counts over years in Urban Dictionary data

A Appendix – Fine-tuning with Urban Dictionary data

A.1 Preprocessing

The full Urban Dictionary data contains $3,534,966$ word definitions. In the dataset provided by Wilson et al. (2020), each entry contains a definition, examples in which the word occurs, number of upvotes & downvotes from website visitors, username of the submitter and a timestamp. As the data is crowd-sourced, many of these entries are noisy and of low quality, and we therefore decided to filter these out and fine-tune RoBERTa only on the best quality definitions. After performing data exploration, we came up with two criteria that we found the most indicative of a definition's quality: the number of upvotes it got, and its upvote/downvote ratio. The distribution of upvotes, downvotes and the upvote/downvote ratios in the dataset can be seen in Figure 6 below. We also note that the number of submissions to Urban Dictionary is relatively well-spread, see Figure 5. This implies that we do not have a strong bias towards more recently popularized slang terms in the dataset, and that we do have representation of the entire time span of interest; $2010 - 2020$.

We keep the entries having more than 20 upvotes and an upvote/downvote ratio of at least 2. This leaves us with $488,010$ Urban Dictionary entries, out of which we randomly sample $100,000$ to reduce the computation time in the fine-tuning process. We use both the definitions and the word usage examples for fine-tuning, producing a final dataset of $200,000$ sequences.

A.2 Training

We randomly split the data into 80% train and 20% test, before training for 10 epochs with an early stopping with patience 3. The batch size was set to

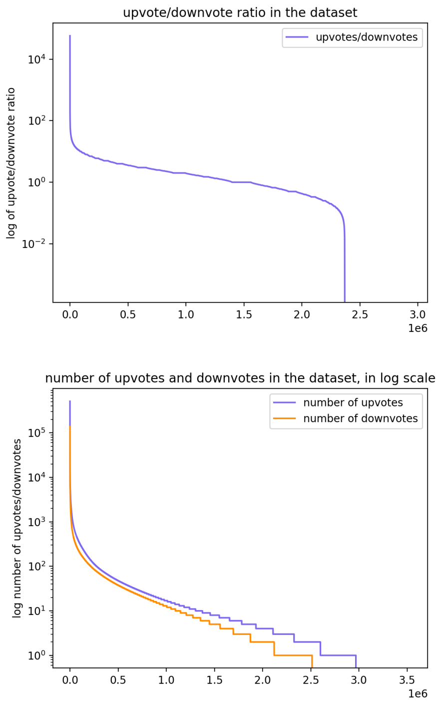


Figure 6: The distributions of (a) upvote/downvote ratio, (b) number of upvotes and number of downvotes among definitions in the dataset in log-scale.

1 in the interest of memory constraints. Following the setup from the pre-training stage as explained in Liu et al. (2019), we use the Adam optimizer (Kingma and Ba, 2015) with $\epsilon = 10^{-6}, \beta_1 = 0.9$ & $\beta_2 = 0.98$ and a linear learning rate decay. For the learning rate, we argue that since the initialized parameters should provide a solution which is already close to the optimum when evaluating on our dataset (our fine-tuning being the very same masked language modeling task as RoBERTa has already been trained on), the learning rate should be smaller. Thus, instead of picking the learning rate $\gamma = 6 \cdot 10^{-4}$ as was done by Liu et al. (2019), we experiment with $\gamma \in \{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$. Training was done using an NVIDIA GeForce GTX 1080 8GB GPU and took around 1 to 1.5 days per model.

B Appendix – Experiments on SemEval-2020

B.1 Distribution-based Metrics

Method: In addition to the distance-based APD metrics, we experiment with two distribution-based ones, namely entropy difference (ED) & Jensen-Shannon Divergence (JSD) (Giulianelli et al., 2020).

We assume a categorical distribution over a set of K_w word senses for word w and time period t. The word sense s_i^w of an occurrence i is then given by:

$$s_i^{wt} \sim Cat(\alpha_1^{wt}, ..., \alpha_{K_w}^{wt}) =: P^{wt}$$

Given two time periods of word sense distributions, we define the ED metric as

$$|H(s^{wt_2}) - H(s^{wt_1})|$$

with entropy $H(\cdot)$. The JSD is given as:

$$\frac{1}{2}KL(P^{wt_1}||M) + \frac{1}{2}KL(P^{wt_2}||M)$$

with $M = \frac{P^{wt_1}+P^{wt_2}}{2}$ and $KL(\cdot||\cdot)$ being the KL-divergence.

We obtain the word sense distributions via a clustering of the representations from both time periods. We experiment with K-Means and Gaussian Mixture Models (GMMs), the latter proposed due to its ability to find more general cluster shapes. We also experiment briefly with Affinity Propagation, which has been used in previous semantic change detection work (Martinc et al., 2020; Kutuzov and Giulianelli, 2020; Montariol et al., 2021). However, we find it to be ill-suited for our purposes since it results in an excessive amount of clusters in comparison to how a human would classify word senses.

For both K-means and GMM, we experiment with selecting the optimal $K_w \in [1, 10]$ through two different procedures. The first one is a slight extension of the method from Giulianelli et al. (2020) – we select the K_w which optimizes the silhouette score (Rousseeuw, 1987) for a set of different initializations. Their approach does not consider the single cluster case however, so we extend it by setting $K_w = 1$ when the best silhouette score is below a threshold of 0.1. For K-Means, we further experiment with an automatic elbow method[5] for

[5] See https://kneed.readthedocs.io/en/stable/index.html

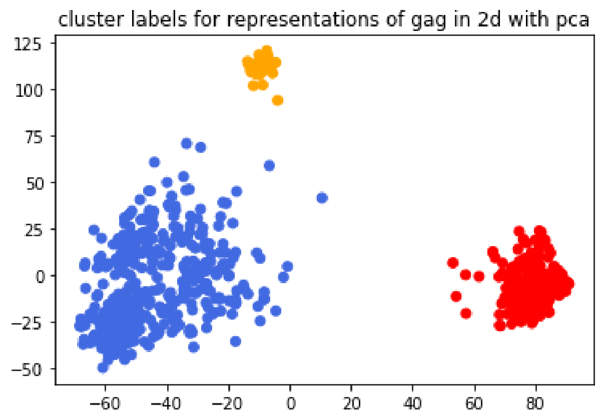


Figure 7: Clusters found with GMM from 2-dimensional PCA representations of the word **gag**.

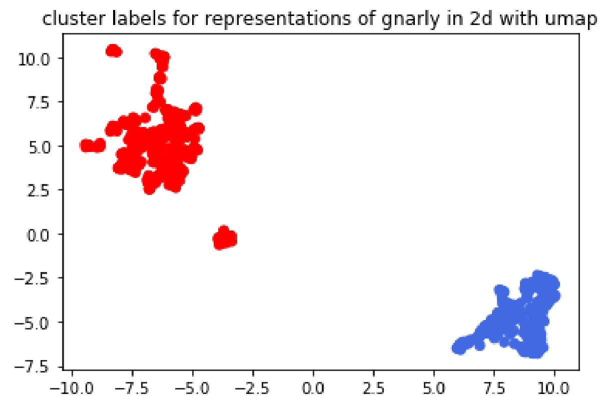


Figure 8: Clusters found with GMM from 2-dimensional UMAP representations of the word **gnarly**.

the sum of squared distances to the cluster centroids, which decreases monotonically with the number of clusters. We again select the cluster assignments with the largest silhouette score for multiple random initalizations. For GMM, we further experiment with taking the model which corresponds to the best Bayesian Information Criterion (Schwarz, 1978).

Clustering examples: In Figure 7 we see three clusters found for "gag." They do not seem to correspond to word senses however: An example from the first cluster is "user i need a pic of you begging if i ' m boiling these because boiled eggs make me gag . :d," an example from the second cluster is "lmao rt user user user so i tried that tuna with cheese and my gag reflexes were in full affect !" and an example from the third cluster is "gag me with a spoon" – all seemingly referring to the sensation of being about to vomit.

We show another example in Figure 8 of the word "gnarly," this time reduced to 2 dimensions using UMAP. Gnarly has three meanings according to the Online Slang Dictionary: It can either mean very good / excellent / cool, gross / disgusting or painful / dangerous. These three word senses are not separated by UMAP and GMM, for instance both "its a good thing one of my roomies is a dude , who else would kill gnarly spiders in my room when i start to hyperventilate" and "rt user bro my wreck on the scooter was so gnarly like it was fun i love shit like that . i wish i could've been on jackass" are put in the first cluster.

Baseline	Score
Combined APD PCA100	0.489
Kutuzov and Giulianelli (2020)	0.605
Kaiser et al. (2020)	0.461
Rother et al. (2020)	0.440

Table 2: Comparison to the three highest performing previous works on the SemEval-2020 Task 1 subtask 2 for the English dataset.

B.2 Variance Explained by PCA components

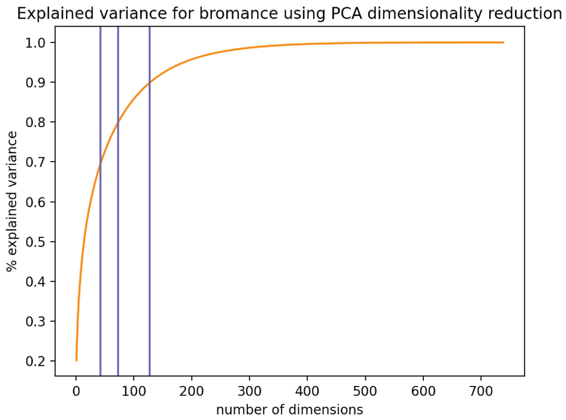
Consider Figure 9 for example plots of how much variance is preserved with PCA on the contextualized representations.

B.3 Results

We further present more results of the experimentation on the SemEval-2020 Task 1 Subtask 2. All tables show the Spearman's rank-order correlation between the change metrics and the ground truths.

In Table 2 we compare our best performing setup to the three best performing previous approaches on SemEval-2020 Task 1 Subtask 2. We see that only Kutuzov and Giulianelli (2020) display a higher score, which might be partially explained by the fact that they fine-tune their model on the SemEval test corpora. We do not do this since our main goal is not to beat state-of-the-art on the shared task, but rather to find a good enough model to detect semantic change in slang.

The results comparing the layer representations can be observed in Table 3. As a side observation we also note that the less isotropic first layer representations seem to perform better than the more isotropic last layer representations.



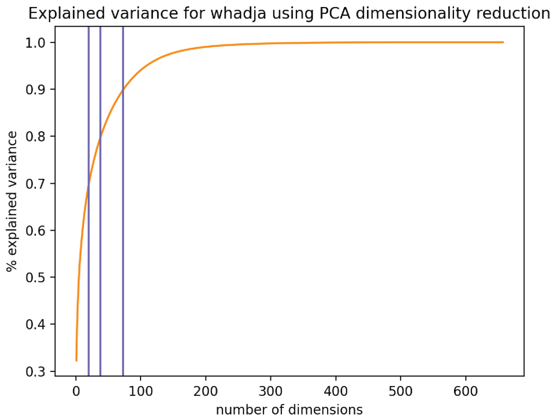


Figure 9: Explained variance by number of components used in PCA for the slang words bromance and *whadja*

In Table 4 we present a comparison across different layer representations for both APD-based and distribution-based metrics. We observe that none of the distribution-based metrics give significant ($p < 0.05$) results, which dimensionality reduction techniques do not manage to improve. While a few of them do have a slight positive correlation, we omit this approach altogether. The APD results on the other hand show a high correlation for many of the configurations, providing an indication of the APD's robustness in detecting semantic change. We show a selection of these in Table 6.

	d_2 **APD**	$d_{\cos}$ **APD**
First layer	0.22	0.234
Last layer	0.07	0.2
Sum of all layers	$\mathbf{0.336}^*$	$\mathbf{0.332}^*$

Table 3: Spearman's rank-order correlation coefficients between our semantic change scores and the ground truth across different layer representations ($p < 0.05$).

Reps	**Clustering**	**Metric**	**Score**	$\boldsymbol{p}$
First	-	APD d_2	0.220	0.190
First	-	APD $d_{\cos}$	0.234	0.164
First	K-Means	ED	-0.079	0.644
First	K-Means	JSD	0.059	0.73
First	GMM	ED	0.051	0.764
First	GMM	JSD	0.072	0.67
Last	-	APD d_2	0.007	0.966
Last	-	APD $d_{\cos}$	0.20	0.236
Last	K-Means	ED	-0.001	0.955
Last	K-Means	JSD	0.202	0.231
Last	GMM	ED	-0.067	0.695
Last	GMM	JSD	-0.096	0.571
All	-	APD d_2	0.336	0.042
All	-	APD $d_{\cos}$	0.332	0.045
All	K-Means	ED	0.033	0.846
All	K-Means	JSD	0.089	0.599
All	GMM	ED	-0.133	0.433
All	GMM	JSD	0.0	0.999

Table 4: Comparison across different layer representations with APDs and distribution metrics, with K_w selected through silhouette scores.

C Appendix – Hybrid Words

We define hybrid words as words that have both a slang and nonslang meaning, i.e. occurring in both Online Slang Dictionary (OSD) and Merriam Webster (MW). In this section, we compare the polysemy, semantic change, frequency shift as well as the absolute frequency change patterns of hybrid words to slang and nonslangs.

Polysemy is collected for hybrid words from OSD and MW separately. Since the MW dictionary may also contain slang meanings, we filter out definitions labeled as slang, informal or vulgar from these scores. The mean polysemy scores of the slang words are (2.074 ± 2.568) and the mean OSD polysemy scores of the hybrid words are (2.580 ± 2.178), with a non-significant difference ($p > 0.05$) in distribution according a permutation test. This tells us that we are not biasing the

APD	Score	p
d_2	0.336	0.042
$d_{\cos}$	0.332	0.045
d_1	0.409	0.012
d_2 and $d_{\cos}$	0.345	0.037
$d_2, d_{\cos}$ and d_1	0.398	0.015

Table 5: Comparison across APD metrics for original representations. Representations are sums across all layers.

Dim	APD	Score	p
PCA2	d_2	-0.153	0.367
UMAP2	$d_{\cos}$	-0.136	0.424
PCA5	$d_{\cos}$	0.209	0.215
PCA5	d_2 and $d_{\cos}$	0.268	0.109
UMAP5	$d_2, d_{\cos}$ and d_1	-0.146	0.39
PCA20	d_2 and $d_{\cos}$	0.42	0.010
PCA50	$d_2, d_{\cos}$ and d_1	0.344	0.037
UMAP50	d_2	-0.158	0.35
PCA100	d_1	0.297	0.074
PCA100	d_2 and $d_{\cos}$	0.489	0.002
UMAP100	$d_{\cos}$	-0.133	0.433

Table 6: Comparison across different dimensions with PCA and UMAP for APD metrics. Representations are sums across all layers.

polysemy scores of the slang words by excluding hybrid words.

As for the nonslang meanings of the hybrid words, we get a mean polysemy score of (6.880 ± 6.080) which is significantly different ($p < 0.001$) from those of the nonslang words (3.079 ± 2.780). This is an interesting observation, implying that had we included nonslang words with hybrid meaning in our nonslang words sample, the difference in polysemy between slang and nonslang words would have been larger. Some example words of this category with high MW polysemy scores include “split”, “down” and “walk”.

For the relative frequency changes, we present the results as histograms in Figure 10. The frequency change in hybrid words seems to fall between those of the slang words and the nonslang words. We observe a mean and standard deviation of -0.154 and 0.608 respectively.

In addition, we compare the absolute relative frequency changes as described in Section 5.3 across slang, nonslang and hybrid words. The histograms are presented in Figure 11. We observe, respectively, a mean and standard deviation of 1.246 & 1.180 for the slang words, 0.950 & 0.724 for the nonslang words and 0.482 & 0.402 for the hybrid words. The difference in mean is significant between the slang and nonslang words ($p < 0.05$), indicating that slang words have undergone a larger absolute change in frequency. Furthermore, we note a highly significant difference ($p < 0.001$) in the mean of the hybrid words compared to both the slang and nonslang word means.

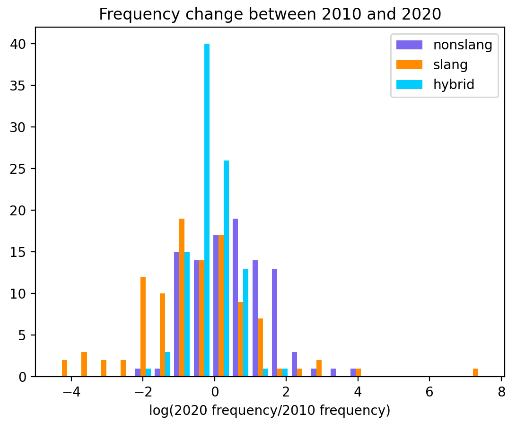


Figure 10: Relative difference in frequency between 2020 and 2010, for slang, nonslang and hybrid words, where a positive score corresponds to an increase in frequency.

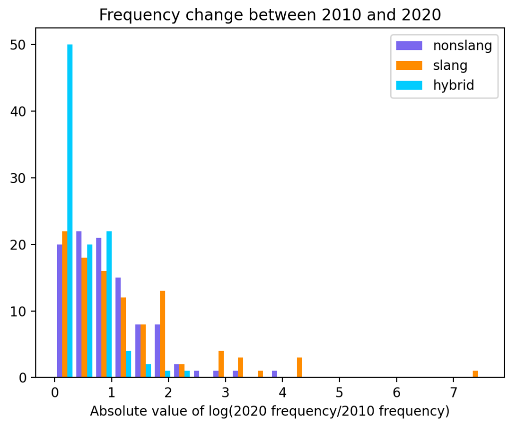


Figure 11: Absolute value of relative difference in frequency between 2020 and 2010, for slang, nonslang and hybrid words, where a larger score corresponds to a larger absolute increase in frequency.

We compare the normalized semantic change scores between the slang, nonslang and hybrid words. Histograms over the semantic change scores are shown in Figure 12. We observe that the distribution over hybrid change scores seem again to be centered between the slang and nonslang dis-

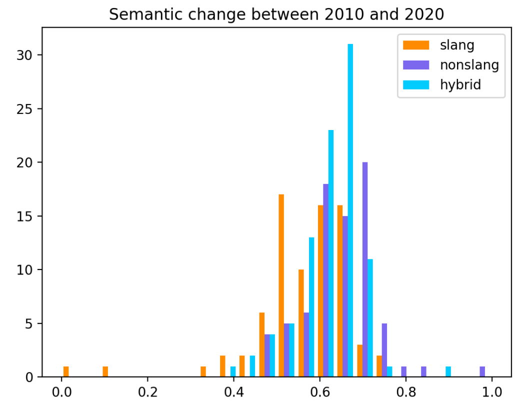


Figure 12: Difference in semantic change score between 2010 and 2020 for slang, nonslang and hybrid words, where a larger score corresponds to a more pronounced semantic change.

tributions, with mean 0.621 ± 0.073. Both the difference in mean compared to slang words and to nonslang words are significant according to permutation tests ($p < 0.001$ for difference to slang words and $p < 0.05$ for difference to nonslang words).

D Appendix – Causal Analysis

D.1 Preliminary on Constraint-based Causal Discovery

Assumptions The constraint-based causal discovery algorithms make use of two main assumptions, namely the global Markov assumption and the faithfulness assumption. The global Markov assumption (Peters et al., 2017) states that all d-separations (defined below) encoded in the causal graph imply conditional independencies in the distribution over the variables contained in the graph. More formally, for a graph $G = (V, E)$ and distribution $\mathbb{P}$ over the variables $\mathbf{X}_V$ it holds that for any disjoint subsets A, B and C of V

$$\mathbf{X}_A \perp_d \mathbf{X}_B | \mathbf{X}_C, \quad \text{in } G$$

$$\Rightarrow \mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_C, \quad \text{in } \mathbb{P}$$

The faithfulness assumption states the converse of the global Markov assumption: All conditional independencies in the distribution are encoded by d-separations in the graph.

d-separation Two nodes $A, B \in V$ are said to be *d-separated* (Geiger et al., 1990) by a set of nodes $Z \subset V$ if for all paths between A and B, at least one of the following holds:

- The path contains a directed chain $A \cdots \rightarrow C \rightarrow \cdots B$ or $A \cdots \leftarrow C \leftarrow \cdots B$ such that $C \in Z$
- The path contains a fork $A \cdots \leftarrow C \rightarrow \cdots B$ such that $C \in Z$
- The path contains a collider $A \cdots \rightarrow C \leftarrow \cdots B$ such that $C \notin Z$ or $C' \notin Z \quad \forall C' \in desc(C)$ (i.e. neither C nor any of its descendants is in Z)

Markov Equivalence Constraint-based algorithms use conditional independency tests in order to identify a *Markov equivalence class* of DAGs. Two DAGs are defined to be Markov equivalent if they have the same skeleton (edges omitting direction) and v-structures. The three vertices A, B and C form a v-structure if $A \rightarrow B \leftarrow C$ and A and C are not directly connected by an edge. Alternatively, two DAGs are Markov equivalent if they describe the same set of d-separation relationships. A Markov equivalence class is the set of all Markov equivalent DAGs.

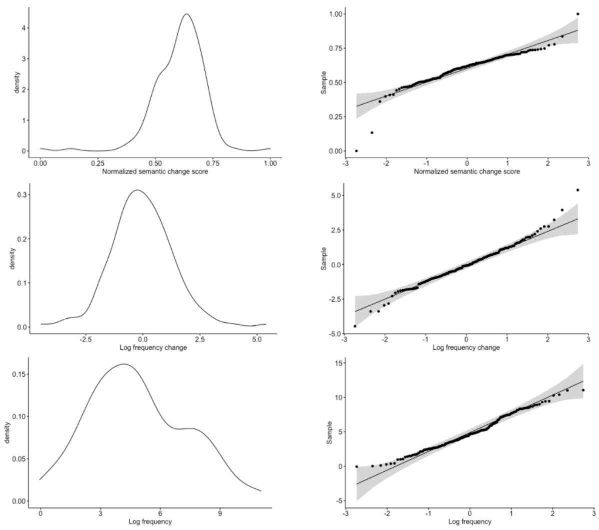


Figure 13: Diagnostic plots for continuous variables, displaying approximate Gaussian shape.

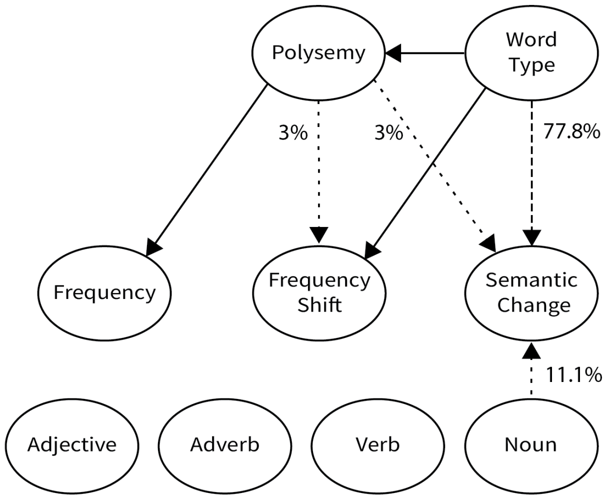


Figure 14: DAG of causal relationships, with the percentage of experiments that found each edge, across different configurations of α and different categorizations of the polysemy score. Solid edges appeared in 100% of the output graphs.

PC Algorithm One common constraint-based algorithm is the PC algorithm (Spirtes et al., 2000). Starting with a full DAG, it eliminates an edge between adjacent vertices i and j if X_i and X_j are conditionally independent given some subset of the remaining variables. This process, including the conditional independence tests, is conducted iteratively starting from a conditioning set of $k = 0$ to $k = |V| - 2$. In addition to the global Markov and faithfulness assumptions, the PC algorithm also assumes causal sufficiency, namely the absence of unobserved confounders. With these assumptions satisfied and access to correct conditional independence relations, it is guaranteed to be sound, complete and uniformly consistent (Kalisch and Bühlmann, 2007).

PC-stable PC-stable is an order-independent extension with the same guarantees as the original (Colombo and Maathuis, 2014).

D.2 Diagnostic Plots

In Figure 13 we present the density and Q-Q plots for semantic change score, log of word frequency and log of frequency change.

D.3 Sensitivity Analysis on Polysemy

Polysemy is a discrete variable which we treat as an ordered factor in the analysis by splitting it into categories. Since polysmey can be plausibly categorized in different ways, we experiment with 9 different categorizations of it and examine the stability of the resulting graphs. For each categorization, we run PC-stable with the three significance levels $\alpha \in \{0.05, 0.03, 0.01\}$. In Figure 14 we present the results of this sensitivity analysis. We see that the edges Between word type and polysemy, from word type to frequency change, as well as the edge from polysemy to frequency, are apparent in all of the configurations. The edge from word type to semantic change is apparent in 21/27 (77.8%) of the configurations. We also observe a few edges very rarely, and therefore label them as noise and do not take them into account for the causal analysis. These consist of an edge from the POS *Noun* to semantic change in 3/27 (11.1%) of the configurations, and edges from polysemy to frequency shift and from polysemy to semantic change each apparent in 1/27 (3.7%) of the configurations.

By inferring the causal graph from a set of categorizations, we make up for the possible noise in the polysemy variable and ensure that the graph isn't sensitive to small variations in the words' polysemy scores.

D.4 Causal Inference

Given the causal DAG in Figure 4, we derive the expression for the average causal effect of word type on semantic change. Define the following random variables: T = word type, X = polysemy, Y = frequency, Z = frequency change and S = semantic change, with respective probability mass functions P_T & P_X and probability density functions f_Y, f_Z & f_S.

Note that $t' \in \{\text{slang}, \text{nonslang}\}$. By the truncated factorization for the connected component of the causal DAG (i.e. excluding POS), we have that

$$\mathbb{P}(s, t, x, y, z | do(T = t')) =$$

$$f_{Y|X}(y|x) f_{Z|T}(z|t) f_{S|T}(s|t) P_{X|T}(x|t) \mathbb{1}_{\{t=t'\}}$$

Marginalizing over T, we get

$$\mathbb{P}(s, x, y, z | do(T = t')) =$$

$$= f_{Y|X}(y|x) f_{Z|T}(z|t') f_{S|T}(s|t') P_{X|T}(x|t')$$

Next, marginalize over the continuous random variables Y and Z to get

$$\mathbb{P}(s, x | do(T = t')) =$$

$$\int_y \int_z f_{Y|X}(y|x) f_{Z|T}(z|t') f_{S|T}(s|t') P_{X|T}(x|t') dz dy =$$

$$\int_y f_{Y|X}(y|x) f_{S|T}(s|t') P_{X|T}(x|t') \underbrace{\left(\int_z f_{Z|T}(z|t') dz \right)}_{=1} dy =$$

$$f_{S|T}(s|t') P_{X|T}(x|t') \underbrace{\int_y f_{Y|X}(y|x) dy}_{=1} =$$

$$f_{S|T}(s|t') P_{X|T}(x|t')$$

Finally

$$\mathbb{P}(s | do(T = t')) =$$

$$\sum_x f_{S|T}(s|t') P_{X|T}(x|t') = f_{S|T}(s|t')$$

Taking the expectation, we get

$$\mathbb{E}[S | do(T = t')] = \mathbb{E}_{S|T}[S|t']$$

E Appendix – Selected Words

In Appendix E we list all the slang and nonslang words used in this study.

Slang	Nonslang
a-list	admitting
badass	adulterous
blankie	agenda
bling	allotted
blowjob	anticlockwise
blumpkin	avoiders
bonehead	awesome
bro	banzai
bromance	bright
bumfuck	butane
bupkis	calorie
chillax	chug
chones	committeeman
colitas	competencies
compo	contenders
conniption	conventionally
crappy	copyediting
dang	deathblow
dis	decomposition
dogg	despoil
duckface	didot
dudette	doubleheader
fanboy	echo
fap	enhancements
gangsta	epilator
glitterati	estimated
gorp	fiddled
gotsta	galavant
gunt	glutton
hasbian	greeting
horribad	grisly
jabroni	groans
jalopy	haircut
jerkwad	heaviest
lame-o	humblest
lemme	ignites
lowkey	inclusive
mcdreamy	intimidator
meme	jugglers
mosey	jute
motherfucking	lawlessness
mozzie	legalist
netizen	milepost
nuker	mistreatment
pedo	moldovan
peeps	morphology
plastered	mushroom
poopy	nonskid
preemie	outlawing
pregos	pantsuit
prettyful	peppy
rapey	performative

Slang	Nonslang
rehab	postural
relly	protocol
roofie	repentant
roshambo	rump
sesh	sabertooth
shart	sailor
shiesty	scallywag
shtick	scheme
sicc	sculptured
sinse	scummiest
skeevy	shield
skyrocket	shylock
slore	snug
snitch	squall
soused	steeple
spam	strap
spec	superabundance
spec-ops	sympathizer
sucky	telogen
tenner	terrifies
thingamabob	trampolining
trisexual	underpainting
tweeker	underrated
twit	unicorn
whadja	unlike
workaround	unmatched
wut	upgrade
zooted	vanadium