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

Anonymous ACL submission

Abstract

Words are not static in their usage and meaning, but evolve over time. An interesting phenomenon in languages is slang, which is an informal language that is considered ephemeral and is often associated with contemporary trends. In this work, we study the semantic change and relative frequency shift of slang words and compare this change with standard, nonslang words. To measure semantic change, we obtain contextualized representations of words, reduce their dimensionality and propose a metric to measure their average pairwise distances between two time periods. We apply causal discovery algorithms and causal inference to uncover the dynamics of language evolution and measure the effect that word type (slang/nonslang) has on both semantic change and frequency shift, as well as its relationship to absolute frequency and polysemy. Our causal analysis shows that slang words undergo less semantic change even though they have larger frequency shifts over time.[1]

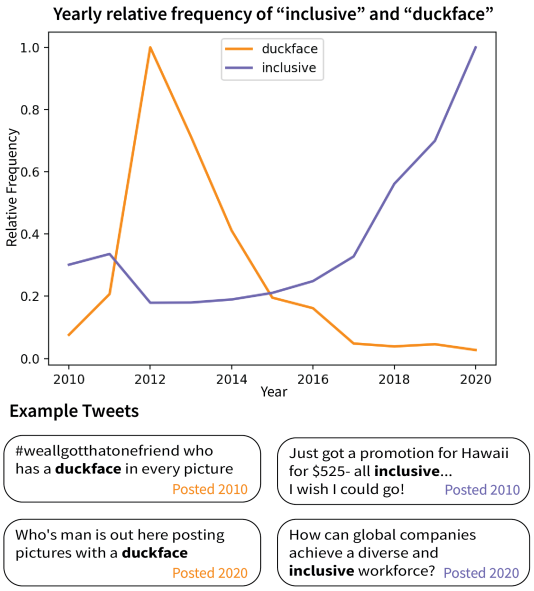

Figure 1: We observe two very different change dynamics for the slang word “duckface” and the nonslang word “inclusive”.

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 frequency under which they are used are not static, but rather evolve over time. Consider for example the word “unicorn”, which in recent years has experienced a metaphorical semantic change (Bloomfield, 1933) from its traditional meaning to also encompass the rare occurrence of startup companies valued at over $1 billion.

In this work, we study semantic change for slang words. Slang is colloquial and informal language commonly associated with particular groups (González, 1998; Bembe and Beukes, 2007), and is characterized by innovation (Mattiello, 2005). Due to these reasons, we believe that slang words follow different change dynamics compared to other words. We investigate the *semantic change*, i.e. the shift in word meaning across time, as well as *frequency shifts* and the *degree of polysemy* of slang words compared to nonslang words.

We formalize this comparison with a causal framework, to establish that there is not just an association, but a direct effect of *word type (slang/nonslang)* on semantic and frequency change. Examining word change dynamics through a causal lens enables us to determine the interactions between the different variables. For example, it allows us to conclude whether word type directly influences semantic change, or rather influences polysemy, which in turn causes semantic change. Our causal analysis follows two steps: (1) finding a causal directed acyclic graph (DAG) (Spirtes et al.,

[1] Our data and code will be released after acceptance.

2000; Pearl, 2009b) to represent the causal relationships between word type, frequency, polysemy and semantic change, and (2) using do-calculus (Pearl, 1995) to evaluate the direct causal effect of word type on semantic change and frequency change. Such an analysis is novel in the semantic change literature, with causal inference being a newly emerging direction of research within the NLP community as a whole (Egami et al., 2018; Keith et al., 2020; Feder et al., 2021; Jin et al., 2021a,b).

We measure both lexical semantic change and frequency change over a 10-year time span by leveraging Twitter data from 2010 and 2020. From this data, we obtain representations using a bidirectional language model (Liu et al., 2019), which we fine-tune on a slang-dense corpus (Wilson et al., 2020). The semantic change score is computed by taking the average pairwise distance (APD) (Sagi et al., 2009; Giulianelli et al., 2020) between dimensionality-reduced representations from the two time points, which we found to be the highest performing method based on experiments on the SemEval-2020 Task 1 benchmark for semantic change (Schlechtweg et al., 2020).

We find that a word being *slang* causes it to undergo more rapid decreases in frequency and slower semantic changes. Moreover, we observe that slang words change faster overall, and polysemy causes words to have higher frequency. To illustrate, consider the slang word "duckface" and the nonslang word "inclusive" as shown in Figure 1. Duckface is the face pose made for photos by pouting the lips (Miller, 2011), often observed in profile pictures during the early 2010s. The semantic meaning of duckface has stayed constant since its inception and it was particularly attached to a trend, after which it mostly disappeared from usage. Indeed, we observe that duckface had a peak in frequency in 2012 and a subsequent rapid decrease, and it furthermore scores very low on semantic change by our model. In contrast, the nonslang word "inclusive" has developed a new usage in recent years, to accommodate people who have historically been excluded (Merriam-Webster, 2019), which is reflected by a high semantic change score, being among the highest in our sample of words.

2 Related Work

2.1 Semantic Change

A typical method for measuring semantic change is by comparing word representations across time periods. For example, Dubossarsky et al. (2016) compare word embeddings obtained using Google n-gram data, to measure semantic change across decades. They find that verbs change faster than nouns, but do not observe an association between word frequency and semantic change. Similarly, Hamilton et al. (2016) measure the cosine distance between word embeddings and discover that polysemous words change at a faster rate, while frequent words change slower. They also note a higher frequency among polysemous words.

These approaches rely on fixed word representations. Limited by assigning one vector to each word, fixed embeddings do not distinguish between multiple word meanings and hence 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 unsupervised neural approaches based on contextualized word embeddings (Peters et al., 2018; Devlin et al., 2018). This has lead to a further stream of work on semantic change detection with contextualized embeddings (Martinc et al., 2020; Kutuzov and Giulianelli, 2020; Montariol et al., 2021; Schlechtweg et al., 2020; Giulianelli et al., 2021).

2.2 Characterization of Slang

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

Mattiello (2005) highlights the innovative nature of slang and the role it plays in enriching the language with neologisms, and claims that it follows unique word formation processes that are different from standard language, such as word clipping and blends. Inspired by this, (Kulkarni and Wang, 2018) propose simple data-driven model for generating slang words according to the processes suggested by Mattiello (2005).

Others have described the ephermality of slang words, which seem to come and disappear from usage more rapidly than standard language (González, 1998; Carter, 2011), however to the best of our

knowledge this has not been previously verified statistically.

2.3 Causal Discovery and Inference

Causality is the study of mining the cause and effect behind data and uncovering how variables influence each other. There are two main tasks in causality: causal discovery, which aims to discover causal relationships, often modeled in the form of a DAG, and causal inference, which concerns determining the effect that intervening on one variable will have on the others.

Causal discovery can be broadly categorized into two main approaches: constraint-based methods and score-based methods. Constraint-based methods rely on conditional independence tests, such as the Peter-Clark (PC) algorithm (Spirtes et al., 2000) as well as its extensions, e.g. the IDA algorithm (Maathuis et al., 2009) and the PC-stable algorithm (Colombo and Maathuis, 2014). Score-based methods, on the other hand, identify the causal graph by optimizing a score function. A representative score-based method is the greedy equivalence search (GES) (Chickering, 2002), which greedily searches over Markov equivalence classes.

The task of causal inference can be facilitated by do-calculus (Pearl, 1995), which estimates causal effects from observational data by establishing the equivalence of interventions and probability distributions estimated from observational data, through conditioning with methods such as backdoor adjustment (Pearl, 1995) and the adjustment criterion (Shpitser et al., 2012).

3 Data Collection

3.1 Slang and Nonslang Word Selection

We select 100 slang words and 100 nonslang words for our study. 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. Since the scope of our study is mainly about single-word expressions, 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", "beef" or "tool". Including words of this type would have created a hardcoded dependency between word type and polysemy, as these words by definition are polysemous. However, since a substantial amount of slang words are hybrid, we perform a separate analysis of these in Appendix C.

As for the reference set of 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).

3.2 Twitter Corpus

To measure the semantic change of slang and nonslang words, we require a dataset that has frequent occurrences of slang and that reflects the general use of colloquial language, and we thus choose the social media platform Twitter as our corpus.

We sample tweets from two different years, 2010 and 2020, which makes it possible to examine the semantic change of words over a 10-year gap. For every slang and nonslang word, and each of the two years, we obtain 200-500 random tweets that contain the word and were posted at one of the over 25 randomly sampled time points within the year. For every tweet, we keep its text, corresponding slang word, tweet ID, and date. As a post-processing step, we remove all duplicate tweets as well as all URLs and hashtags from the tweets. To protect user privacy, we replace all user name handles with the generic word "user."

We obtain 170,135 tweets in total. On average we have 370 slang and 333 nonslang tweets per word from 2010, and 323 slang and 254 nonslang tweets per word from 2020.

4 Collecting Causal Variables

We explore the potential causal effect of a word's type (slang/nonslang) on its semantic change. We additionally test the hypothesis that slang words appear and dissipate faster, and if so, whether this is due to a causal effect. For these purposes, we 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 4.1)
- **Frequency Change:** The relative difference in frequency the word has undergone between 2010 and 2020 in the Twitter corpus (Section 4.2)

[2] http://onlineslangdictionary.com/

Year	Slang	Nonslang
2010	1,931	1,507
2020	13,560	8,802
Overall Increase	×7.0	×5.8

Table 1: Average daily counts for both slang and nonslang words, in 2010 and 2020

- **Polysemy:** The number of senses a word has (Section 4.3)
- **Semantic change:** The semantic change score of the word from 2010 to 2020 in the Twitter corpus (Section 4.4)

4.1 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 time points 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, and then averaged for the causal analysis. Due to the growing popularity of social media, the number of tweets has significantly increased throughout the decade. Therefore, we divide the tweet 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 average daily counts in both years can be seen in Table 1. This normalization factor is also justified by the fact that both the number of active daily users on Twitter and the number of tweets per day had approximately a 7-fold increase over this time (GDELT Project, 2019; Internet Live Stats; Dean, 2021).

4.2 Frequency Change

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

$$\text{FreqChange}(w) = \log \frac{x_{2020}(w)}{x_{2010}(w)} \qquad (1)$$

where, $x_k(w)$ is the frequency of word w in year k. This was proven 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 $-0.48(\pm 1.65)$ for slang words and $0.53(\pm 1.07)$ for nonslang words, where a negative score corresponds to a decrease 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. Namely, the highest increase in frequency was for the slang word "incel", which went from being tweeted an average of 0.28 times a day in 2010, to being tweeted about 3400 times a day in 2020, whereas the steepest decreases in frequency was for the slang word "celebutante".

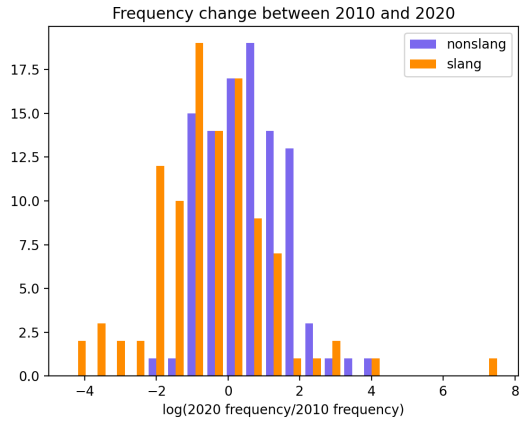

Figure 2: Relative difference in frequency between 2020 and 2010, for slang and nonslang words, where a positive score corresponds to an increase in frequency. We see more slang words show a decrease in frequency than nonslang ones.

We also examine the absolute value of the change in frequency in equation (1) 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 absolute frequency than nonslang words ($p < 0.05$). See Appendix C for more details.

4.3 Polysemy

We define a word's polysemy score as the number of distinct senses it has. This definition also encapsulates potential cases of homonymy; we choose not to make a distinction between polysemy and homonymy in this analysis. For nonslang words, we take the number of senses the word has in WordNet (Fellbaum, 1998; Princeton University, 2010), which was designed to group words into distinct senses. If a word does not appear in the WordNet database, we take the number of distinct definitions it has on the Merriam Webster dictionary. For slang words, polysemy score is determined by their number of definitions on the Online Slang Dictionary.

We note a decreasing empirical probability mass function for the polysemy score, with mean 2.49 and standard deviation 3.21. More polysemous words appear to also be more frequent in our dataset – the log transform of average frequency and polysemy display a highly significant ($p < 0.001$) linear correlation coefficient of 0.34.

4.4 Semantic Change Score

In this section we explain in detail how we obtain the semantic change scores. We start by fine-tuning a bi-directional language model on a slang-dense corpus (Section 4.4.1), after which we survey the literature and propose metrics (Section 4.4.2) that we use to perform an extensive experimentation study to find the most suitable one (Section 4.4.3). Finally, we apply this metric to our sets of slang and nonslang words (Section 4.4.4).

4.4.1 Obtaining Contextualized Representations

As input for semantic change scoring, we leverage the contextualized representations obtained from a bi-directional language model. As a first step, we familiarize the 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). 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. 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), which is based on Transformers (Vaswani et al., 2017) and pre-trained on the following datasets: BookCorpus (Zhu et al., 2015), CC-News (Nagel, 2016), OpenWebText (Gokaslan and Cohen, 2019) and Stories (Trinh and Le, 2018) – all presumably limited in the use of slang. Beyond performance gains compared to the original BERT (Devlin et al., 2018), we select this model since it uses byte-pair encoding with bytes instead of characters as sub-units, allowing 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 version RoBERTa base for computational reasons.

We train the model using the Adam optimizer (Kingma and Ba, 2017) with learning rates $\gamma \in \{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$. For semantic change scoring we proceed to use the model with the lowest loss on the test set, which is the one trained with a learning rate $\gamma = 10^{-6}$. For more details on training configurations, we refer to Appendix A.2.

4.4.2 Quantifying Semantic Change

In order to find 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.[3] Our space of configurations include layer representations, dimensionality reduction techniques and semantic change metrics. In addition to the average pairwise distance (APD) metrics, we also experiment with distribution-based metrics (see Appendix B.1).

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, both when conditioning on words and more surprisingly, when considering all words. However, the level of isotropy may vary according to the layer from which the representations are retrieved – it has been shown that the word self-similarity of representations in BERT decreases with the layer index (Ethayarajh, 2019; Cai et al., 2021), which would imply that deeper/higher layers have a higher degree of isotropy. We hypothesize that a more isotropic space lends itself better to semantic change detection, since the distance metrics will be more pronounced. This leads us to experiment with three different representations of our fine-tuned RoBERTa model: taking only the first layer, only the last layer or summing all layers.

[3] We do keep in mind 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. It might therefore be that our model successfully detects change in modern informal language, but fails to perform well on the SemEval task.

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

Table 2: Spearman's rank-order correlation coefficients between our semantic change scores and the ground truth across different layer representations ($p < 0.05$).

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

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). UMAP works by reducing a high-dimensional graph to maintain local as well as global structure. While UMAP has been known to be able to find nicely separated clusters (Coenen et al., 2019), the Euclidean distances in the reduced space are very sensitive to hyperparameters and it does not retain an interpretable notion of absolute distances. Thus, UMAP is not suitable for pure distance-based metrics like APD. We therefore also experiment with PCA, which in contrast finds and projects the data onto the directions with the largest variances.

APD Metrics for Semantic Change: Given word representations $\mathcal{X}_t = \{\boldsymbol{x}_{1,t}, ..., \boldsymbol{x}_{n_t,t}\}$ for time period t we define the APD between representations of two periods as

$$\text{APD}(\mathcal{X}_{t_1}, \mathcal{X}_{t_2}) = \frac{1}{n_{t_1} n_{t_2}} \sum_{\substack{\boldsymbol{x}_{i,t_1} \in \mathcal{X}_{t_1} \\ \boldsymbol{x}_{j,t_2} \in \mathcal{X}_{t_2}}} d(\boldsymbol{x}_{i,t_1}, \boldsymbol{x}_{j,t_2}) \,, \tag{2}$$

for some distance metric $d(\cdot, \cdot)$. 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

$$||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2^2 = ||\boldsymbol{x}_1||_2^2 - 2\boldsymbol{x}_1^T \boldsymbol{x}_2 + ||\boldsymbol{x}_2||_2^2 \tag{3}$$

$$\leq ||\boldsymbol{x}_1||_2^2 + ||\boldsymbol{x}_2||_2^2 \tag{4}$$

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

$$d_{2,\cos}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{0.5 \cdot d_2(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sqrt{||\boldsymbol{x}_1||^2 + ||\boldsymbol{x}_2||^2}} + \frac{d_{\cos}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{4} \tag{5}$$

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

4.4.3 Evaluating the Semantic Change Scores

We first present the results of three types of layer representations for Euclidean and Cosine APD metrics. The results can be observed in Table 2. We see that summing all layer representations give the highest correlation with the true change scores. Consequentially, we only present the results using these representations henceforth. 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.

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 4.4.2 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 3, showing that both combining the metrics and PCA dimensionality reduction improve the performance. More results and comparisons to baselines are presented in Appendix B.3. From these we observe 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 show in Appendix B.2. In addition, both norm-based metrics perform worse with dimensionality reduction.

4.4.4 Semantic Change Scores on the Twitter Dataset

For evaluating semantic change on the Twitter dataset we choose the best performing configuration on SemEval, which is the Euclidean and cosine

combined APD metric computed on the sum of all layer representations, being reduced to 100 dimensions with PCA. This is further justified seeing as the combined APD metric performs best across all dimensions except $h = 2$ and the dimension size of $h = 100$ performs well across all APD metrics.

For the semantic change scores, we use words that have more than 150 tweets in each time period after the filtering step described in Section 3.2, in order to ensure that we get meaningful representations. This leaves us with 80 slang and 81 nonslang words. The resulting semantic change scores are shown in Figure 3. The mean semantic change scores are $0.731(\pm 0.011)$ for slang words and $0.739(\pm 0.009)$ for nonslang words.

Some of the slang words with the lowest semantic change scores were "whadja" (0.674), "dudette" (0.710) and "duckface" (0.714), while the slang words "skyrocket" (0.746) and "dogg" (0.749) displayed a relatively high semantic change. Among the nonslang words, "anticlockwise" (0.774) and "inclusive" (0.752) undergo a large change, and the lowest scores are displayed by "terrifies" (0.720) and "underpainting" (0.721).

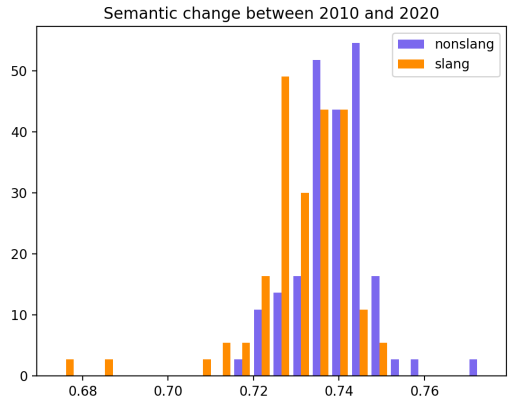

Figure 3: Difference in semantic change score between 2010 and 2020 for slang and nonslang words, where a larger score corresponds to a more pronounced semantic change.

5 Causal Analysis

Previous works (Dubossarsky et al., 2016; Hamilton et al., 2016) have suggested causal factors in the context of semantic change, but none have however applied a causal framework to analyze and confirm these relationships. Here, we inspect the underlying mechanisms of semantic change with causal discovery methods, which we use to infer the effect that word type has on semantic change and frequency shift.

5.1 Causal Discovery

We refer the reader to Appendix D.1 for a short preliminary on causal discovery. For learning the causal graph, we choose the constraint-based algorithm PC-stable (Colombo and Maathuis, 2014), which is an order-independent variant of the original PC algorithm (Spirtes et al., 2000). It evaluates causal links through conditional independence tests, which should be chosen according to the underlying data distribution. Since we are learning a mixed graphical model (Lauritzen, 1996; Lee and Hastie, 2015), consisting of both continuous and categorical data, this calls for tailoring the tests to each specific set of variables we are considering. 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 score, relative frequency change and log of word frequency are then all assumed to be approximated well by a Gaussian distribution, which is confirmed by diagnostic density and Q-Q plots.

As for the the ordinal polysemy variable, we discretize and treat it as a categorical variable, by splitting it into three categories: one word sense (monosemous), 2-5 word senses, or more than five word senses. We also check for robustness with respect to different categorizations, see Appendix D.2. Word type is categorical in nature. For the two 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 $\alpha \in \{0.01, 0.03, 0.05\}$.

5.2 Resulting Causal Structure

In Figure 4 we see the result from the above approach, using a significance level of $\alpha = 0.03$ or $\alpha = 0.05$ for the conditional independence tests, both of which resulted in similar results across configurations. See Appendix D.2 for a sensitivity

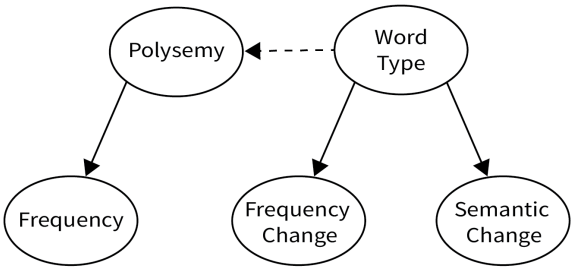

Figure 4: DAG representing the causal relationships in our dataset.

analysis.

We first observe that word type has a direct causal effect on both the semantic change score and the frequency change, without any confounders. Between word type and polysemy we observe a weak dependence, as this edge was not present in all of our aforementioned configurations and could furthermore not be oriented by the PC-stable algorithm. We manually orient the edge as outgoing from type and ingoing to polysemy however, 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 weakly confounded by word type. In the case of absolute frequency, it is independent of semantic change ($p > 0.05$) both when conditioning on polysemy and/or word type, as well as in the empty conditioning set. The faithfulness assumption would suggest the latter to be a dependency, which highlights the uncertainty of the link from word type to polysemy.

5.3 Causal Effect of Word Type on Semantic Change and Frequency Change

We evaluate the average causal effect of word type T on semantic change S as:

$$\mathbb{E}[S|do(T = \text{nonslang})] - \mathbb{E}[S|do(T = \text{slang})] \quad (6)$$

In our case there are no confounders, as presented in Figure 4, and this equation therefore reduces to the difference between the conditional distributions:

$$\mathbb{E}[S|T = \text{nonslang}] - \mathbb{E}[S|T = \text{slang}] \quad (7)$$

See Appendix D.3 for a derivation. The case of frequency change is analogous.

We estimate the expectations by the sample means and get an average causal effect of 0.008, which is a highly significant value ($p < 0.001$) based on a permutation test (Edgington, 1969).

For the observed changes in relative frequency, calculated according to Equation 1, we record an average causal effect of 1.017 which is highly significant ($p < 0.001$) via a permutation test.

6 Discussion

We analyze the dynamics of frequency 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 in particular are more likely to greatly decrease in frequency.

To ensure that this effect is the result of a direct causal effect, and not mediated through another variable or subject to confounders, we model the data with a causal graph, by also considering potential interacting variables such as a word's polysemy and average absolute frequency. We discover that there is no influence of confounders, nor are there mediators between a word's type (slang/nonslang) and its semantic change or its frequency change, which confirms a direct causal effect.

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 (Feltgen et al., 2017; Kroch, 1989), as well as a known positive correlation between polysemy and frequency (Casas et al., 2019; Lee, 1990). However, we do not find a causal effect of polysemy or absolute frequency on semantic change, in contrast to suggestions made in previous works (Hamilton et al., 2016).

7 Conclusion

In this paper, we analyze the change dynamics of slang, a unique and informal part of language, and compare it to that of standard, nonslang words. We do so by applying a combined APD metric to contextualized representations obtained from Twitter data, and further use causal discovery to model the factors that influence word change dynamics. We discover a causal relationship between a word being slang and having slower semantic change, as well as more rapid decreases in frequency.

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.

Leonard Bloomfield. 1933. *Language*. New York: Allen Unwin.

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.

David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554.

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.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *CoRR*, abs/1906.02715.

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

Brian Dean. 2021. How many people use twitter in 2021? Accessed Oct 8 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Haim Dubossarsky, D. Weinshall, and E. Grossman. 2016. Verbs change more than nouns: a bottom-up computational approach to semantic change.

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.

Eugene S. Edgington. 1969. Approximate randomization tests. *The Journal of Psychology*, 72(2):143–149.

D. I. Edwards. 2000. *Introduction to Graphical Modelling*, 2nd edition. Springer.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. How to make causal inferences using texts.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

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.

GDELT Project. 2019. Visualizing eight years of twitter's evolution: 2012-2019. Accessed Oct 8 2021.

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.

Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarova. 2021. Grammatical profiling for semantic change detection.

Aaron Gokaslan and Vanya Cohen. 2019. Open webtext corpus.

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.

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.

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.

Internet Live Stats. Twitter usage statistics. Accessed Oct 8 2021.

Zhijing Jin, Zeyu Peng, Tejas Vaidhya, Bernhard Schoelkopf, and Rada Mihalcea. 2021a. Mining the cause of political decision-making from social media: A case study of COVID-19 policies across the US states. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics.

Zhijing Jin, Julius von Kuegelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021b. Causal direction of data collection matters: Implications of causal and anticausal learning in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics.

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. *CoRR*, abs/2008.03164.

Katherine A. Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

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

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. Uiouva at semeval-2020 task 1: Contextualised embeddings for lexical semantic change detection. *CoRR*, abs/2005.00050.

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. *CoRR*, abs/1907.11692.

Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. 2009. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164.

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 o fslang in standard and non-standard english.

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.

Sebastian Nagel. 2016. Cc-news.

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

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

Judea Pearl. 2009b. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press.

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. *CoRR*, abs/1802.05365.

Princeton University. 2010. About wordnet. Accessed Oct 4 2021.

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, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection.

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

Ilya Shpitser, Tyler VanderWeele, and James M. Robins. 2012. On the validity of covariate adjustment for estimating causal effects.

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.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724.

A Appendix – Fine-tuning with Urban Dictionary data

A.1 Preprocessing

The full Urban Dictionary data contains $3,534,966$ word definitions. As the data is crowd-sourced, many of the definitions are noisy and of low quality. Moreover, as fine-tuning RoBERTa is an expensive task, we decided to filter out most of the definitions and fine-tune the model only on the best quality ones. 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; $2010 - 2020$.

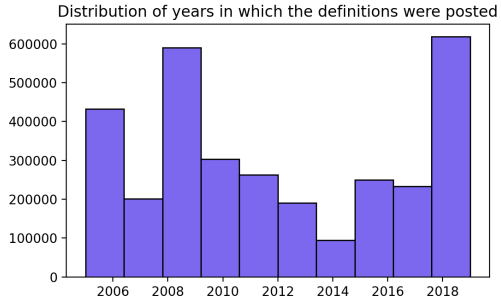

Figure 5: Frequency counts over years in Urban Dictionary data

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 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

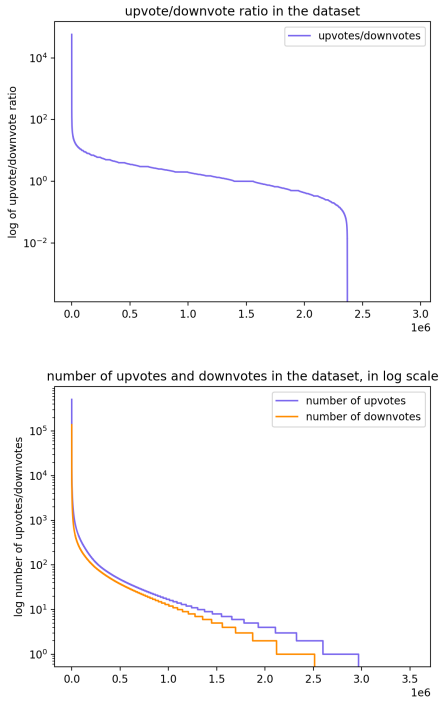

Figure 6: The distributions of (a) upvote/downvote ratio, (b) number of upvotes and number of downvotes among definitions in the dataset in log-scale.

(Kingma and Ba, 2017) 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 $\gamma \in \{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$.

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[4] for 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.

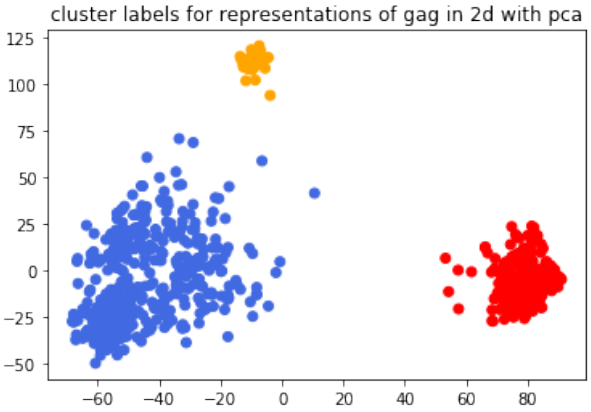

Figure 7: Clusters found with GMM from 2-dimensional PCA representations of the word **gag**.

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.

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

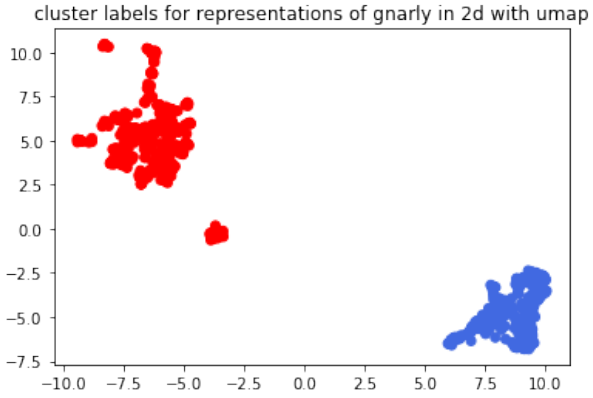

Figure 8: Clusters found with GMM from 2-dimensional UMAP representations of the word **gnarly**.

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 4 we compare our best performing setup to the three best performing previous approaches on SemEval-2020 Task 1 Subtask 2.

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

In Table 2 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 7.

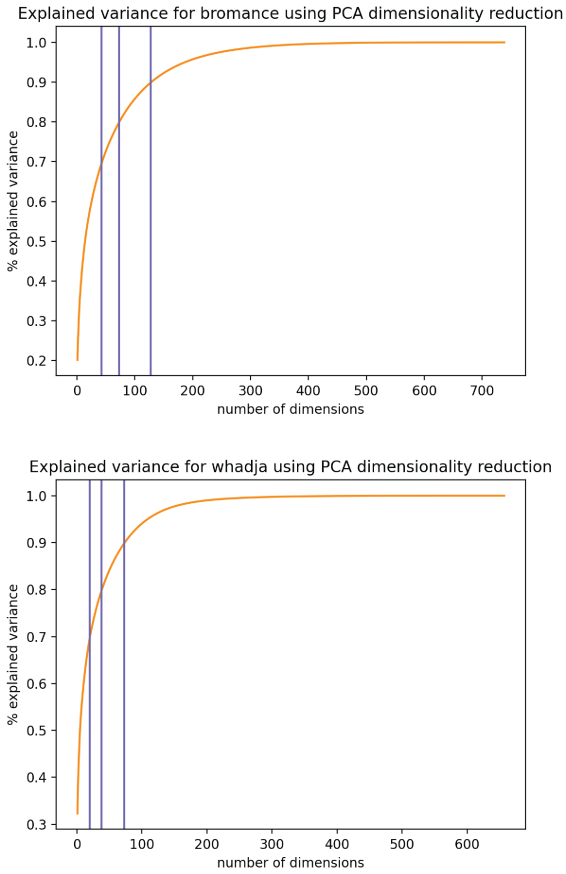

Figure 9: Explained variance by number of components used in PCA for the slang words bromance and *whadja*

C Appendix – Hybrid Words and Absolute Change

We compare the frequency and semantic change distributions of hybrid words, which we define to be words that have both slang and nonslang meanings, to those of exclusively slang and nonslang words.

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 4.2 across slang, nonslang and hybrid words. The histograms are presented in Figure 11. We observe, respectively, a mean and standard deviation of 1.246 & 1.18 for the slang words, 0.950 & 0.724 for the nonslang words and 0.482 & 0.402 for the hybrid

Reps	Clustering	Metric	Score	p
First	-	APD d_2	0.22	0.19
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 $d_{\cos}$	0.2	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 5: Comparison across different layer representations with APDs and distribution metrics, with K_w selected through silhouette scores.

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 6: 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	0.01
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 7: Comparison across different dimensions with PCA and UMAP for APD metrics. Representations are sums across all layers.

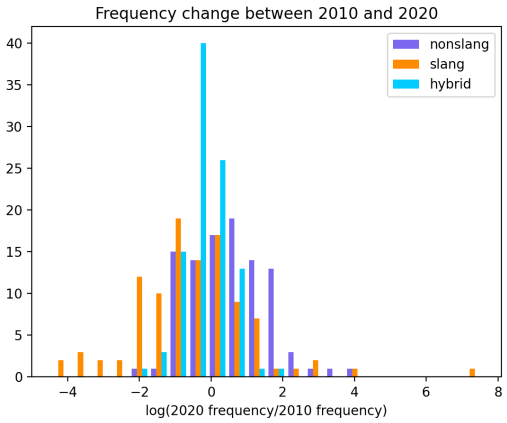

Figure 10: Relative difference in frequency between 2020 and 2010, for slang, nonslang and hybrid words, where a positive score corresponds to an increase in frequency.

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.

For the semantic change scores, 92 hybrid words remain after the filtering step described in Section 3.2. 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 distributions, with mean and standard deviation of 0.736 and 0.0074 respectively. 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).

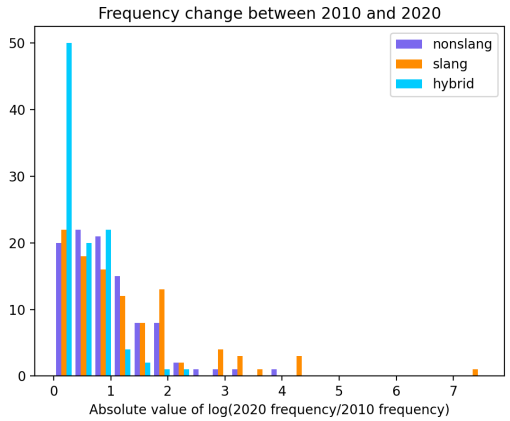

Figure 11: Absolute value of relative difference in frequency between 2020 and 2010, for slang, nonslang and hybrid words, where a larger score corresponds to a larger absolute increase in frequency.

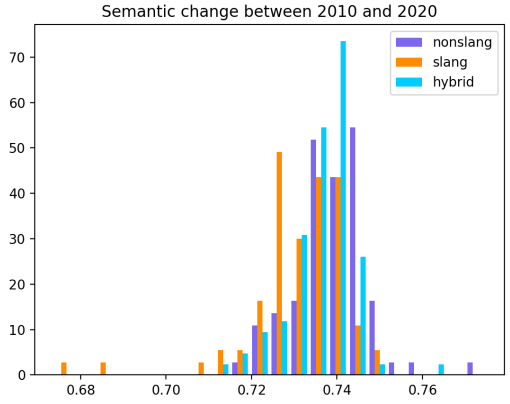

Figure 12: Difference in semantic change score between 2010 and 2020 for slang, nonslang and hybrid words, where a larger score corresponds to a more pronounced semantic change.

D Appendix – Causal Analysis

D.1 Preliminary on Causal Discovery

The constraint-based causal discovery algorithms make use of two main assumptions, namely the global Markov property and the faithfulness assumption. The global Markov property (Peters et al., 2017) states that all d-separations (Geiger et al., 1990) 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 is defined as the opposite direction: All conditional independencies in the distribution are encoded by d-separations in the graph. Constraint-based algorithms use conditional independency tests and can, under certain assumptions, identify a Markov equivalence class of directed acyclic graphs that fulfill both conditions. Two DAGs are defined to be Markov equivalent if they have the same skeleton (edges omitting direction) and v-structures. The three nodes A, B and C form a v-structre if $A \rightarrow B \leftarrow C$ and A and C are not directly connected by an edge.

D.2 Causal Discovery Sensitivity

In Figure 13 we present the results of our sensitivity analysis for the causal discovery with PC-stable. For each significance level, we apply ten different categorizations for the polysemy variable. Stratifying by test significance level ($\alpha = 0.05, \alpha = 0.03, \alpha = 0.01$), the edge appearances for word type to polysemy were $80\%, 70\%$ and 0%, for polysemy to frequency change $20\%, 10\%$ and 0% and for polysemy to semantic change $30\%, 20\%$ and 0%. We therefore discard the causal links from polysemy to semantic change and frequency change, and label the link between word type and polysemy as "weak".

D.3 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 (Pearl, 2009a) for the causal DAG, 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

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

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

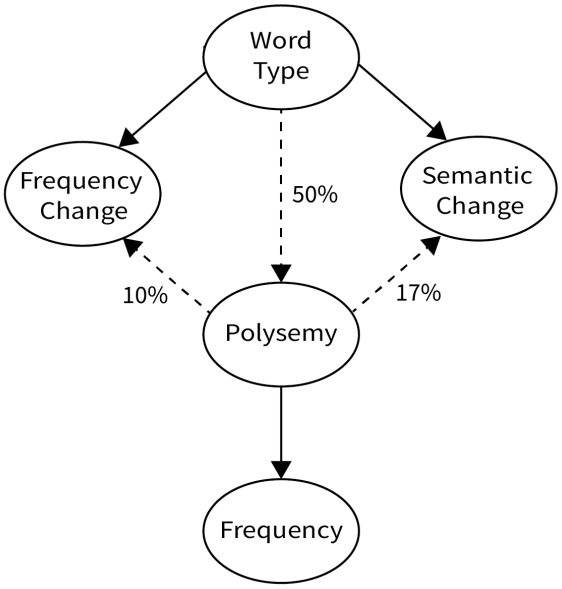

Figure 13: DAG of causal relationships, with the percentage of experiments that found each edge, across different configurations of α and different categorizations of polysemy score. Solid edges appeared in 100% of the output graphs.

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

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