Predicting Long-Term Citations from Short-Term Linguistic Influence

Sandeep Soni and **David Bamman**
University of California, Berkeley
{sandeepsoni,dbamman}@berkeley.edu

Jacob Eisenstein
Google Research
jeisenstein@google.com

Abstract

A standard measure of the influence of a research paper is the number of times it is cited. However, papers may be cited for many reasons, and citation count offers limited information about the extent to which a paper affected the content of subsequent publications. We therefore propose a novel method to quantify linguistic influence in timestamped document collections. There are two main steps: first, identify lexical and semantic changes using contextual embeddings and word frequencies; second, aggregate information about these changes into per-document influence scores by estimating a high-dimensional Hawkes process with a low-rank parameter matrix. We show that this measure of linguistic influence is predictive of *future* citations: the estimate of linguistic influence from the two years after a paper's publication is correlated with and predictive of its citation count in the following three years. This is demonstrated using an online evaluation with incremental temporal training/test splits, in comparison with a strong baseline that includes predictors for initial citation counts, topics, and lexical features.

1 Introduction

The citation count of a paper is a standard, easily measurable proxy for its influence (Cronin, 2005). Researchers have shown that citation count is strongly correlated with the quality of scientific work (e.g., Lawani, 1986), the recognition that a paper or an author gets (e.g., Inhaber and Przednowek, 1976), or in policy decisions such as assessment of scientific performance (e.g., Cronin, 2005). Consequently, citation count is a ubiquitously deployed and important measure of a paper with whole subfields of research dedicated to its analysis (Bornmann and Daniel, 2008).

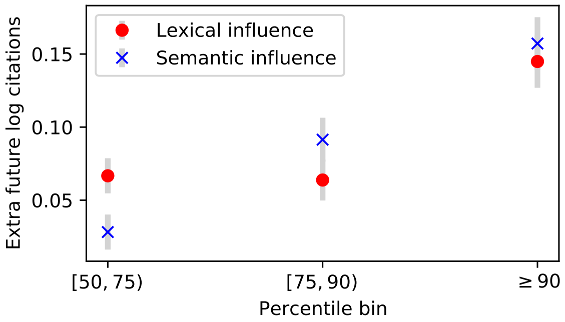

Figure 1: Research papers that are more linguistically influential within an initial time window tend to receive more citations in the long term. The x-axis shows lexical and semantic influence, binned into quantiles (see § 2); the y-axis shows the corresponding regression coefficients and standard errors, in units of Z-normalized log future citations (see § 5.3). To give a sense of scale, for papers published in 2012, being in the top decile of semantic influence corresponds to an 14.5% increase in long-term citations, as compared to control-matched papers that received the same number of short-term citations and covered similar topics but were in the bottom half by semantic influence.

However, papers may be cited (or not cited) for many reasons, and citation count alone is insufficient to explain the emergence and the spread of research ideas and trends. For this reason, we turn to content analysis: to what extent can the text of a research paper be said to influence the trajectory of the research community? In this paper, we present a novel technique for estimating the influence of documents in a timestamped corpus. To demonstrate the validity of the resulting measure of linguistic influence, we show that it is predictive of *future* citations. Specifically, we find that: (1) papers that our metric judges as highly influential in the short term tend to receive more citations in the long term; (2) short-term linguistic influence increases the ability to predict long-term citations over strong baselines.

Our modeling approach focuses on semantic changes, and treats the temporal usage of semantic innovations as emissions from a parametric low-rank Hawkes process (Hawkes, 1971). The parameters of the Hawkes process correspond to the linguistic influence of each paper, aggregated over thousands of linguistic changes. The changes themselves are identified through analysis of contextual embeddings, with the goal of finding words whose meaning has shifted over time (Traugott and Dasher, 2001). Though there are several computational methods to detect semantic changes (e.g., Kim et al., 2014; Hamilton et al., 2016; Rosenfeld and Erk, 2018; Dubossarsky et al., 2019), including methods based on contextual embeddings (e.g., Kutuzov and Giulianelli, 2020), our proposed method focuses on detecting smooth, non-bursty semantic changes; we also go further than other methods by distinguishing old and contemporary usages of an identified semantic change.

We show through a multivariate regression that our estimates of semantic influence of each paper are positively correlated with their long-term citations, even after controlling for the initial citations, the content of the paper in terms of topics, and the lexical influence of the paper (see Figure 1). Further, we formulate long-term citation prediction as an online prediction task, constructing test sets for successive years. The addition of semantic influence as features to a model once again improves the predictive performance of the model over baselines. In summary, our contributions are as follows:[1]

- We empirically demonstrate a link between long-term citation count and short-term linguistic influence, using both regression analysis (§ 5.3) and an online prediction task (§ 5.4).
- We present a method to estimate semantic influence using a parametric Hawkes process (§ 2.1). To achieve this, we find semantic changes and convert the usage of each change into a cascade (§ 2.2). We also show that the method can be applied to quantify lexical influence.
- We present a method to identify monotonic semantic changes from timestamped text using contextual embeddings (see § 2.2.1).

2 Methodology

This section describes our method for estimating the linguistic influence of each document in a timestamped collection. Our work builds on the theory of point process models (Daley et al., 2003), in which the basic unit of data is a set of marked event timestamps. In our case, the events correspond to the use of an innovative word or usage; the mark corresponds to the document in which word or usage appears. To estimate linguistic influence of individual documents, we fit a parametric model in which per-document influence parameters explain the density of events in subsequent documents. We first describe the modeling framework in which these influence parameters are estimated (§ 2.1) and then describe how event cascades are constructed (§ 2.2) from semantic changes (§ 2.2.1) and lexical innovations (§ 2.2.2).

2.1 Estimating document influence from timestamped events

A marked cascade is a set of marked events $\{e_1, e_2, \ldots, e_N\}$, in which each event $e_i = (t_i, p_i)$ corresponds to a tuple of a timestamp t_i and a mark p_i. Assume a set of marked cascades, indexed by $w \in \mathcal{W}$, with each mark belonging to a finite set that is shared across all cascades. In our application, each cascade corresponds to the appearances of an individual word or word sense, and each mark is the identity of the document in which the word or word sense appears.

Point process models define probability distributions over cascades. In an inhomogeneous point process, the distribution of the count of events between any two timestamps (t_1, t_2) is governed by the integral of an intensity function $\lambda(t, w)$. A Hawkes process is a special case in which the intensity function is the sum of terms associated with previous events (Hawkes, 1971). We choose the following special form,

$$\lambda(t, w) = c_w + \sum_{i:t_i^{(w)} < t} \alpha_{p_i^{(w)}} \kappa(t - t_i^{(w)}), \qquad (1)$$

where κ is a time-decay kernel such as the exponential kernel $\kappa(\Delta t) = e^{-\gamma \Delta t}$ and c_w is a constant. The parameter of interest is α, which quantifies the influence exerted by the document $p_i^{(w)}$ on subsequent events.[2]

[1] The code and relevant data from our paper can be found at `http://github.com/sandeepsoni/contextual-leadership`

[2] In the more general multivariate Hawkes process, the intensity function can depend on the identity of "receiver" of influence. This enables the estimation of pairwise excitation parameters $\alpha_{i,j}$, as in the work of Lukasik et al. (2016), to give an example from the NLP literature. However, it would be difficult to estimate pairwise excitation between thousands of documents, as required by our setting.

Our application focuses on research papers, which historically have been published in a few bursts — at conferences and in journals — rather than continuously over time. For this reason we simplify our setting further, discretizing the timestamps by year. The evidence to be explained is now of the form $n(t, w)$, the count of word or sense w in year t. We model this count as a Poisson random variable, and estimate the parameters c_w and α by maximum likelihood.

2.2 Building event cascades

To estimate the parameters in Equation 1, we require a set of timestamped events. Ideally these events should correspond to evidence of linguistic innovation. We consider two sources of events: semantic innovations (here focusing on words whose meaning changes over time) and lexical innovations (words whose usage rate increases dramatically over time).

We now introduce some notation used in the remainder of this section. Let a document be a sequence of discrete tokens from a finite vocabulary $\mathcal{V}$, so that document i is denoted $X_i = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n_i)}]$, with n_i indicating the length of document i. A corpus is similarly defined as a set of N documents, $\mathcal{X} = \{X_1, X_2, ..., X_N\}$, with each document associated with a discrete time $t_i \in \mathcal{T}$.

2.2.1 Using contextual embeddings to identify semantic changes

We use contextual embeddings to identify words whose meaning changes over time, following prior work on computational historical linguistics (e.g., Kutuzov and Giulianelli, 2020, see § 6 for a more comprehensive review). A contextual embedding $\boldsymbol{h}_i^{(k)} \in \mathbb{R}^D$ is a vector representation of token k in document i, computed from a model such as BERT (Devlin et al., 2019). When the distribution over $\boldsymbol{h}$ for a given word changes over time, this is taken as evidence for a change in the word's meaning.

Let $m_{t-,w}$ and $m_{t+,w}$ be the count of the word w up to and after time t, respectively. Specifically,

$$m_{t-,w} = \sum_{i:t_i \leq t} \sum_{k}^{n_i} \mathbb{1}(x_i^{(k)} = w)$$
$$m_{t+,w} = \sum_{i:t_i > t} \sum_{k}^{n_i} \mathbb{1}(x_i^{(k)} = w)$$

Average representations of the word w up to and after time t, respectively, are calculated as follows.

$$\boldsymbol{v}_{t-,w} = \frac{1}{m_{t-,w}} \sum_{i:t_i \leq t} \sum_{k}^{n_i} \boldsymbol{h}_i^{(k)} \mathbb{1}(x_i^{(k)} = w)$$
$$\boldsymbol{v}_{t+,w} = \frac{1}{m_{t+,w}} \sum_{i:t_i > t} \sum_{k}^{n_i} \boldsymbol{h}_i^{(k)} \mathbb{1}(x_i^{(k)} = w)$$

Further, the variance in the contextual embeddings of the word w over the entire corpus is calculated by taking the variance of each component of the embedding,

$$\mathbf{s}_w = \frac{1}{m_w} \sum_{i,k:x_i^{(k)}=w} \left(\boldsymbol{h}_i^{(k)} - \boldsymbol{\mu}_w\right)^2, \qquad (2)$$

with μ_w equal to the mean contextualized embedding of word w.

A semantic change score for a word w for a time t is then the variance-weighted squared norm of the difference between its average pre-t and post-t contextualized embeddings (also known as the squared Mahalanobis distance):

$$r(w, t) = (\boldsymbol{v}_{t-,w} - \boldsymbol{v}_{t+,w})^\top \boldsymbol{S}_w^{-1} (\boldsymbol{v}_{t-,w} - \boldsymbol{v}_{t+,w}), \qquad (3)$$

with $\boldsymbol{S}_w = \text{Diag}(\mathbf{s}_w)$.

Correction for frequency effects. Both the mean and variance are estimated with larger samples for timestamps in the middle of $\mathcal{T}$ in comparison to the initial and final timestamps. Consequently, the distance metric suffers from high sample variance for values of t near these endpoints. The discrepancy is corrected by replacing the diagonal covariance $\boldsymbol{S}_w$ in Equation 3 with an alternative covariance $\tilde{\boldsymbol{S}}_w$ that reflects that additional uncertainty due to sample size. Specifically, we approximate the standard error of the mean v_{t-} as $\sqrt{\boldsymbol{S}/m_{t-}}$, and analogously for v_{t+}. Then $\tilde{\boldsymbol{S}}_w$ is defined as the product of these two approximate standard errors,

$$\tilde{\boldsymbol{S}}_w = \sqrt{\frac{\boldsymbol{S}_w}{m_{t^-,w}}}\sqrt{\frac{\boldsymbol{S}_w}{m_{t^+,w}}} = \frac{\boldsymbol{S}_w}{\sqrt{m_{t^-,w} m_{t^+,w}}}. \tag{4}$$

Finally, $t^* = \operatorname{argmax}_t r(w,t)$ is selected as the transition point for the change in meaning of w. The changes are identified by sorting the words by $\max r(w,t)$ and applying a set of basic filters explained in § 4. To give some intuition:

- If w changes in meaning at time t, then the difference in its representation up to t and after t should be high. The metric in Equation 3 captures this precisely by calculating the term $\boldsymbol{v}_{t^-,w} - \boldsymbol{v}_{t^+,w}$.

- Difference in average embeddings can be high for seasonal or bursty changes seen in words such as *turkey* which is referred to the bird more frequently at the time of American holidays (Shoemark et al., 2019). Rescaling the difference by the inverse variance encourages detection of monotonic changes.

- For rare words, the mean embeddings will be less reliable. The $\sqrt{m}$ terms in $\tilde{\boldsymbol{S}}$ have the effect of emphasizing high-frequency words for which changes in the mean embedding are likely to be significant.

Distinguishing old and new usages. The previous step yields semantic innovations and their transition time. Simply identifying semantic changes is insufficient, since at any given time a word could be used in its old or new sense with respect to its time of transition. To categorize every usage of a semantic innovation w, the contextual embeddings are passed through a logistic regression classifier that predicts whether the usage is before or after the transition time. At the end of this step a sequence of embeddings for any semantic innovation is converted to a sequence of binary labels denoting their usage. For each word w, the cascade $(e_1^{(w)}, e_2^{(w)}, \ldots e_{N_w}^{(w)})$ is formed by filtering the usages to those that are classified as corresponding to the newer sense, with each event $e_i^{(w)}$ containing a timestamp $t_i^{(w)}$ and a document identifier $p_i^{(w)}$. These cascades are the evidence from which we estimate the per-document *semantic* influence scores α_s, as described in § 2.1.

Why contextual embeddings? Embeddings provide a powerful tool for understanding language change, offering more linguistic granularity than measures of change in the strength or composition of latent topics (e.g., Griffiths and Steyvers, 2004; Gerow et al., 2018). Prior work has employed diachronic non-contextual embeddings (e.g., Soni et al., 2021b). Such methods require each word to have a single shared embedding in each time period. During periods in which a word is used in multiple senses, the non-contextual embedding must average across these senses, making it harder to detect changes in progress.

2.2.2 Identifying lexical changes

Unlike semantic changes, whose identification requires representations such as contextual embeddings, lexical changes are identified simply by comparing frequency changes. Specifically, for every word in a vocabulary we vary the segmentation year, say t, for the word and calculate the relative frequency up to and after t. We then take the best relative frequency ratio across the years as the score of lexical change for that word and aggregate to form a list of changes by sorting on this score. In contrast to semantic changes, all the usages of lexical changes are used to form cascades. These cascades are the evidence from which we estimate the per-document *lexical* influence scores α_ℓ, again using the methods in § 2.1.

2.3 Overview

To summarize the method for computing semantic influence:

1. Compute the score $r(w,t)$ for each word w and time t as described in Equation 3 (with the adjusted covariance term from Equation 4), and threshold to identify semantic changes.

2. For each word selected in the previous step, classify each usage as either “old” or “new”, and build a cascade from the timestamps of the new usages.

3. Aggregating over all the cascades, estimate the influence parameters α_i for each document in the collection.

A visual summary of the entire methodological pipeline is given in Figure 2.

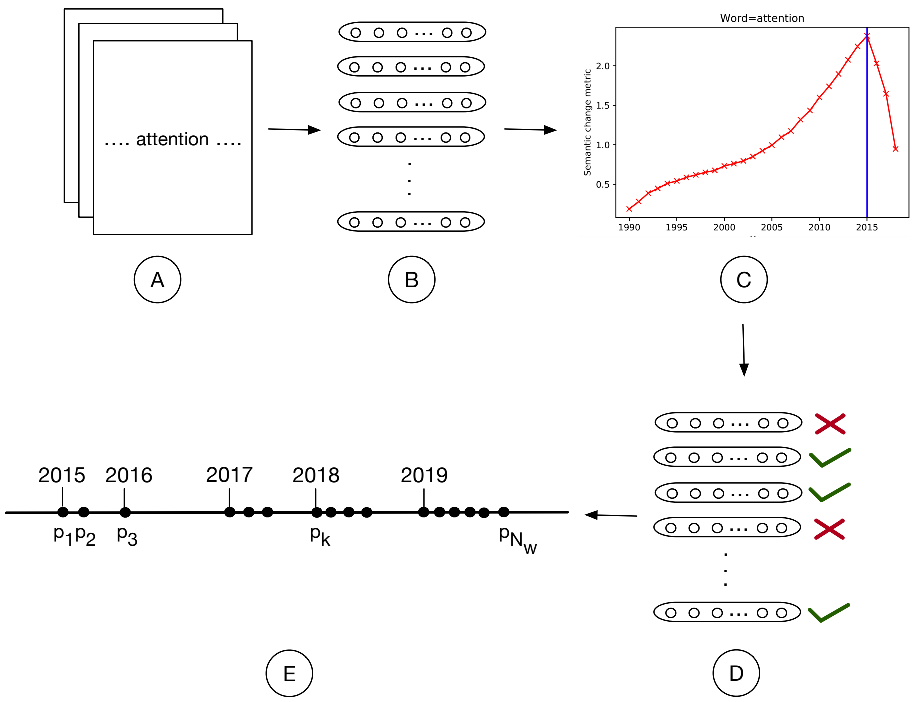

Figure 2: **Methodological pipeline**. The steps in our method can be summarized as follows for an example word *attention*. **(A)** depicts a collection of research papers that mention *attention*; **(B)** is a collection of contextual embeddings for *attention* across the entire corpus; **(C)** uses the contextual embeddings to find the transition point and the magnitude of the change; **(D)** uses the contextual embeddings to classify usages as old (marked with red crosses) or new (marked with green ticks) with respect to the transition time; **(E)** is a depiction of the event cascades comprising of timestamp and paper_id (p_i) pairs.

3 Data

To construct a collection of research papers, we focus on papers that are included in the ACL anthology. We collected the ACL anthology bibliography file[3] and converted the bib entries from the file as `JSON` objects; we retained the title of the paper, the year in which it was published, and the venue.

We then stripped all whitespace and special characters from the title of the paper. These stripped titles and the year of publication are matched with papers in *s2orc* corpus (Lo et al., 2020)[4]. Matched papers that have a valid pdf parse (i.e full text of the paper) are retained. Though the s2orc dataset contains papers from as far back as 1965, the coverage in the early years is sparse with few or no papers in many of the early years. As a result, the data is further filtered to retain only the papers that appear from 1990 to 2019 ($\mathcal{T} = [1990, 2019]$). Descriptive statistics of the curated corpus is given in Table 1.

Number of papers	36645
Years	1990–2019
Mean (median) cites per paper	6.68 (1)
Mean (median) words per paper	3291 (3248)

Table 1: **Dataset summary.** Descriptive summary of the curated ACL corpus from *s2orc* dataset.

4 Experimental Setup

For this study, multilingual BERT is used as the contextualizing model even though our data is English papers. This is to handle even those English language papers that have foreign language tokens. Specifically, the `bert-base-multilingual-uncased` model from the Hugging face (Wolf et al., 2020) library is used.[5] The size of the contextualized embeddings is 3144 dimensions after concatenating the final four layers.

Continued pretraining Previous work has shown that the quality of the contextual embed-

[3] Downloaded from `https://aclanthology.org/` on 12/9/2021

[4] 20200705v1 version

[5] `https://huggingface.co/bert-base-multilingual-uncased`

dings improves when the pretrained BERT is further trained on domain-specific text (e.g., Gururangan et al., 2020). For this study, we continued to pretrain BERT model for 3 epochs to optimize the masked language modeling objective. The probability of masking is set to 15 %.

Wordpiece aggregation Since BERT learns subword embeddings by breaking tokens into wordpieces, the embeddings of the wordpieces need to be aggregated to get a representation of a token. This aggregation is done by taking the average of the wordpiece embeddings.[6]

Data preprocessing Non-English papers in the corpus are ignored from the analysis by identifying the language of the papers using `langid` (Lui and Baldwin, 2012). The vocabulary $\mathcal{V}$ is constructed by retaining words that appear at least 10 times in the abstracts and do not appear in more than 90 % abstracts. Each paper is first segmented by whitespace and then broken into chunks of 200 tokens. Only alphabetic tokens are retained.

Classifying individual usages of semantic innovations The off-the-shelf logistic regression classifier from `scikit-learn` is used to mark every individual instance of a semantic innovation as new or old. To avoid overfitting, we use l_2 regularization; all other inputs to the classifier are set to default. 4-fold cross-validation is performed to get the final assignment of labels from the classifier.

Word filters We keep words in our vocabulary if they are composed only of alphabetic characters, occur in at most 90% of the papers, and occur a minimum of 30 times in the entire corpus. We also eliminate words whose length is less than or equal to 2 characters.

Estimation To estimate the parameters of the Hawkes process, we use `scipy.optimize`, which internally uses the `L-BFGS` solver.

5 Results

5.1 Semantic changes

We identified 2910 semantic changes that capture several technical concepts in language research. The top changes and the period in which their meanings shift are shown in Table 2.

The evolution of language research, from the earlier focus on syntax and sequence processing using latent variable models to the current paradigm of using deep learning, is neatly summarized by the semantic innovations that the method identifies. Changes such as *tokenization* and *transducers* from the late nineties are indicative of the then-structural approach to core NLP research.

The earlier part of the 2000s saw changes in terms such as *plan* (see Table 5 for context in which the term appears), whose narrow usage in messaging applications broadened to other applications. The next decade also saw changes in terms such as *kernel* and *probabilistic*. These indicate the methodological changes that were underway during this period, with NLP research being dominated by a mix of kernel and bayesian methods during this decade (e.g., Moschitti, 2004; Blei et al., 2003). Methodological innovations such as conditional random fields (Lafferty et al., 2001) and the rise of domain adaptation (e.g., Chelba and Acero, 2004; Daumé III, 2007) are also evidenced by terms such as *conditional* and *adaptation*.

With the rise of neural approaches, words such as *representations*, *network*, and *decoder* underwent semantic changes between the years 2013 to 2017. Another prominent example of this shift is the term *attention*, shown in Figure 3, which shifts from its standard, broad usage to the more technical and focused usage with respect to neural networks around 2015.

5.2 Lexical changes

We selected the top 3000 lexical innovations to approximately match the number of semantic innovations. The lexical changes capture the introduction and rise in popularity of terms in language research. Unlike semantic changes, lexical changes are identified only by their change in frequency.

Among the top changes are terms such as *bert*, *lstm*, *adam*, and *mturk* which are examples of new models, algorithms, tools, and technology introduced in language research. On the other hand, example changes such as *factuality* (e.g., Saurí and Pustejovsky, 2012; de Marneffe et al., 2012; Soni et al., 2014) and *sarcasm* (Riloff et al., 2013; Ptáček et al., 2014) indicate the rise in popularity of these concepts during specific years.

Abbreviations such as *sts* and *mt*, and names of languages such as *de* and *indonesian* are two categories of changes that prominently feature among

[6] Elementwise max as an alternative strategy of aggregation was also tried and performed similar in detecting changes.

Period	Semantic Changes
2000-2002	*system*, *data*, *plan*, *language*, *sentence*
2003-2005	*state*, *task*, *relation*, *development*, *shared*
2006-2008	*event*, *topic*, *comments*, *points*, *side*
2009-2011	*media*, *user*, *social*, *users*, *neural*
2012-2014	*network*, *hidden*, *embedding*, *layer*, *representations*
2015-2017	*attention*, *representation*, *sequence*, *mechanism*, *decoder*
2018-2020	*self*, *heads*, *glue*, *contextualized*, *attacks*

Table 2: **Examples of semantic changes.** We show a few handpicked examples amongst the top semantic changes in different periods. More context is shown in Table 5.

top lexical changes. While the former indicates the necessity of naming technical concepts with memorable shortform, the latter is indicative of the rise in multilingual language research.

5.3 Regression analysis

Our objective is to test whether the linguistic influence of a paper is positively correlated with its rate of future citations. However, many factors can confound our analysis including, but not limited to, the early citations a paper gets and the content of the paper. To control for these confounds and test our hypothesis, we frame the problem as a multivariate regression where features that proxy linguistic influence are incorporated alongside proxy features of other factors to predict the future citations. For our analysis in this section and § 5.4, we consider papers published in or after the year 2000, since the density of innovations appearing in these years is higher. The total number of papers in this interval is 19153.

Our unit in the multivariate regression is a research paper and the dependent variable is the Z-normalized logarithm of its future citations. The Z-normalization uses a unique mean and variance for each year of publication, which helps to account for secular trends in the overall rate of citation over time. By "future citations" we mean the difference between the number of citations a paper gets five years after its publication (hereon referred as "long-term citations") and the number of citations a paper gets two years after its publication (hereon referred as "short-term citations"). For example, for a paper published in 2012, the short-term citations are from the period $2012 - 2014$ and the long-term citations are the citations accrued between $2015 - 2017$.

To test the impact of semantic influence, we include three baseline regression models. In the first baseline, M1, we include the Z-normalized short-term citations and a constant term as our only covariates. M2, our second baseline, consists of all covariates in M1 and the topic distribution of a paper learned from an LDA model (Blei et al., 2003). The topic distribution is taken as a coarse representation of the content of the paper. Our final baseline is M3 which contains all the covariates from M2 in addition to categorical covariates corresponding to quantiles of the Z-normalized *lexical* influence, α_l, of each paper. We consider four quantiles: $< 50^{th}$ percentile, $\geq 50^{th}$ and $< 75^{th}$ percentile, $\geq 75^{th}$ percentile and $< 90^{th}$ percentile, and $\geq 90^{th}$ percentile. Finally, our experimental model, M4, has all the covariates from M3 and additional categorical covariates corresponding to the quantiles of the Z-normalized *semantic* influence, α_s, of each paper. The quantiles are divided in the same way as lexical influence.

The experimental model can be compared with the baseline models by their goodness-of-fit, measured by the log-likelihood of the data; analogously, the null hypothesis is that the goodness-of-fit of the experimental model is no better than that of the baseline models. Statistically, the likelihood ratio, our test statistic, follows a χ^2 distribution with the excess number of parameters in the experimental model as the degrees of freedom. The null hypothesis can be rejected if the observed test statistic is determined to be unlikely under this distribution.

The regression coefficients are shown in Table 3.[7] Not surprisingly, short term citations are the strongest predictor of long-term citations, as seen by the strength of the regression coefficient. The regressions further reveal a strong relationship between semantic influence and long-term citations: M4 obtains a significantly improved fit over M3, our strongest baseline ($\chi^2(3) = 91, p \approx 0.0$). Without additional controls, the average rate of long-term citations for the top quantile of semantic influence is 3 times the long-term citation rate for

[7] Due to space limitations we omit the topic coefficients from the table. The topics and their coefficients for M4 are shown in the appendix in Table 6.

Predictors	M1	M2	M3	M4
Constant	-0.000 (0.005)	-0.080 (0.036)	-0.106 (0.036)	-0.116 (0.036)
Initial Citations	0.763 (0.005)	0.740 (0.005)	0.727 (0.005)	0.718 (0.005)
Lex. Inf. Q2			0.079 (0.012)	0.067 (0.012)
Lex. Inf. Q3			0.086 (0.014)	0.064 (0.014)
Lex. Inf. Q4			0.181 (0.017)	0.145 (0.018)
Sem. Inf. Q2				0.028 (0.012)
Sem. Inf. Q3				0.091 (0.015)
Sem. Inf. Q4				0.157 (0.018)
Log Lik.	-18828	-18681	-18615	-18569

Table 3: **Regression analysis**. We show the results of long-term citations for various ablations. Each column indicates a model, each row indicates a predictor, and each cell contains the coefficient and, in parentheses, its standard error. Topics are included as controls in models M2-4, but for clarity their coefficients are reserved for the supplementary material. Results for the best bandwidth parameter (γ=100), selected by the best heldout log-likelihood, are produced here whereas the regression results for other bandwidth settings are in the supplementary material.

the bottom quantile. With additional controls, the top quantile of semantic influence amounts to an increase in the expected citations by a factor of 1.2, in comparison to the papers in the bottom quantile.

5.4 Predicting future citations

We now turn to predicting the long-term citations from semantic influence and the other predictors described in § 5.3. To more closely match the scenario of true future prediction, we formulate this as an online prediction task, in which the model is trained on past data to make predictions about future events (Karimi et al., 2015; Søgaard et al., 2021). Formally, to make predictions about papers published in year t, we use information from the interval $[t, t+2]$ to compute the predictors: short-term citations, lexical influence, and semantic influence. We then make predictions about citations in years $[t+3, t+5]$. To estimate the weights of these predictors, we assume access to training data up to year $t+2$. We then increment t and make predictions about the papers published in the next year. In this way, all papers published in the period 2001-2014 appear in one of the test folds.

The rest of the setup is similar to § 5.3 except one important difference. For the prediction task, we plug in estimates of lexical and semantic influence for all the values of $\gamma = \{0.001, 0.01, 0.1, 1.0, 10.0, 100.0\}$ as predictors in the model. The results of the online prediction of long-term citations are shown in Table 4. The performance is measured using mean squared error (MSE) between the predicted and ground-truth values. The model M4, which includes our measure of semantic influence, achieves the lowest error in 13 of 14 years, and it gives a more accurate prediction than M3 for 57.8% of the 18554 papers in this slice of the dataset.

6 Related Work

6.1 Linguistic change and influence

Several computational methods have been developed to identify changes in language (Eisenstein, 2019). Of particular interest are techniques for detecting semantic changes in a text corpus. Such techniques are based on a range of representations, including frequency statistics (e.g., Bybee, 2007), static, type-level word embeddings (e.g., Sagi et al., 2009; Wijaya and Yeniterzi, 2011; Kulkarni et al., 2015; Hamilton et al., 2016), and contextual word embeddings (e.g., Kutuzov and Giulianelli, 2020; Giulianelli et al., 2020; Montariol et al., 2021). Here, we use contextual embeddings which are, in principle, advantageous over static embeddings

Publication Year	M1	M2	M3	M4
2001	0.739	0.737	0.732	0.731
2002	0.759	0.757	0.755	0.754
2003	0.681	0.679	0.673	0.674
2004	0.623	0.622	0.613	0.606
2005	0.57	0.568	0.554	0.54
2006	0.583	0.581	0.565	0.548
2007	0.504	0.501	0.501	0.486
2008	0.517	0.515	0.506	0.491
2009	0.481	0.479	0.475	0.473
2010	0.516	0.516	0.508	0.497
2011	0.49	0.489	0.482	0.476
2012	0.525	0.524	0.519	0.511
2013	0.511	0.51	0.505	0.498
2014	0.445	0.444	0.431	0.423
All Years	0.529	0.528	0.52	0.511

Table 4: **Online predictive analysis** We show the performance in terms of MSE for the ablated models on the online citation prediction task. The first column indicates the publication year, the subsequent columns are the various ablations as seen in Table 3, and each cell shows the MSE. The last row is the micro-averaged MSE over all the examples. Note that smaller values indicate better predictive performance.

as they can distinguish the dynamics of co-existing senses.

Although there are many methods to detect changes, only a few computational studies find leaders or followers of these changes, which is important in order to understand who carries influence. By modeling lexical changes as cascades on a network, researchers have inferred that they propagate because of influence from strong ties (e.g., Goel et al., 2016). Other researchers have identified leaders and followers of individual semantic changes and aggregated them to induce a leadership network between the sources (Soni et al., 2021a). Our work shares similarities with these prior studies but is distinct: We use similar cascade modeling techniques but for semantic changes, which are considerably harder to construct.

Most relevant to our current work is that of Soni et al. (2021b) who find that semantically progressive scientific research papers get more citations. Semantic progressiveness — a measure of linguistic novelty — is calculated by comparing the old meaning of semantic innovations with their contemporary meaning in the context of the document. Our current work is different from this prior work in a key aspect: We estimate and establish a link between citation influence and semantic influence, instead of semantic novelty.

6.2 Citation influence

Citation count has historically been used as a proxy for the influence of a scientific article (Fortunato et al., 2018), of researchers (Börner et al., 2004), and is shown to be strongly correlated with scientific prestige (Cole and Cole, 1968).Relevant to our work are studies that establish a link between citation influence to different measures of linguistic progressiveness. Kelly et al. (2018) find that progressiveness as measured in terms of difference in textual similarity between old and new patents is predictive of future citations of a patent. Similarly, Soni et al. (2021b) find that progressiveness measured as the early adoption of words of with newer meanings is predictive of citations of a paper. In contrast, in this paper, we find a link between linguistic influence in the short term to the future citations of the paper.

7 Conclusion

We have presented a new technique for quantifying semantic influence in time-stamped documents. Quantitative analysis demonstrates that this measure of semantic influence is strongly correlated with long-term citations a paper receives, and leads to improvement in the prediction of future citations. Our tool offers additional granularity in terms of linguistic influence, which can supplement structural measures of influence based on citation counts. Though we present quantitative analyses for scholarly documents in computational linguistics, our tool could be applied to scholarly documents in other research areas or to documents such as patents or court opinions where citation counts are considered structural measures of influence. We plan to focus on these applications in the future.

8 Limitations

A simplifying assumption in this paper is there exists one dominant sense of a change before and after the transition point. This assumption may not hold for every change, in general, but helps in developing computational methods to identify a large array of changes. In future work, we plan to extend the ability of our method to identify coevolving senses.

A fundamental limitation of the Hawkes Process is the closed-world assumption that all events are attributable to other observed events. This limitation is particularly relevant to our setting, where we observe only papers published in the ACL anthology, but those papers influence and are influenced by a much wider discourse, which includes not only other academic research papers but also software artifacts, books, and social media. In practice, this means that our method might wrongly assign credit to "fast follower" papers that are the first to adopt ideas published outside the ACL universe. Similarly, we make no attempt to measure the extent to which ACL anthology papers influence writing that is published elsewhere.

More generally, we cannot show whether the relationship between linguistic influence and citations is causal. The temporal asymmetry ensures that the future citations are not themselves causes of linguistic influence, but we cannot exclude the possibility that there is a common cause for both phenomena. For example, it seems likely that factors such as the overall quality of the research and the fame of the authors both contribute to the extent to which a paper drives the adoption of linguistic features in the short term, and to the number of citations it receives in the long term. Our regression analysis includes control variables for some potential common causes, such as topics, but it is not possible to control for all other potential confounders. Hence, our analysis should be considered *correlational* and *not causal*. Future work could focus on establishing and quantifying a causal link between linguistic influence and citations.

9 Ethics Statement

This paper offers a new tool for understanding scientific communication. Because this tool quantifies the linguistic impact of research papers, there is the possibility that it could be used for consequential decisions such as hiring, promotion, and funding. This implies a "leaderboard" approach to scholarship that would overvalue the most fashionable mainstream research topics, while penalizing research that has a deep impact in a relatively small community. Similar concerns have been raised about other measures of academic impact: Jorge Hirsch, the inventor of the H-index, noted that his metric could have "severe unintended negative consequences," and urged evaluators to go beyond any single index to consider the broader context when considering an individual's scientific contributions (Conroy, 2020). The same applies to semantic influence metric defined in this paper.

Acknowledgements

We thank David Mimno and William Cohen who gave feedback on the initial draft of this paper. We also thank all the anonymous reviewers for their time and suggestions which led to improvements of the paper. The research reported in this article was supported by funding from the National Science Foundation (IIS-1813470 and IIS-1942591).

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Katy Börner, Jeegar T Maru, and Robert L Goldstone. 2004. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5266–5273.

Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*.

Joan L Bybee. 2007. Diachronic linguistics. In *The Oxford handbook of cognitive linguistics*.

Ciprian Chelba and Alex Acero. 2004. Adaptation of maximum entropy capitalizer: Little data can help a lo. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 285–292, Barcelona, Spain. Association for Computational Linguistics.

Stephen Cole and Jonathan R Cole. 1968. Visibility and the structural bases of awareness of scientific research. *American sociological review*, pages 397–413.

Gemma Conroy. 2020. What's wrong with the h-index, according to its inventor. *Nature Index*, 24.

Blaise Cronin. 2005. *The hand of science: Academic writing and its rewards*. Scarecrow press.

Daryl J Daley, David Vere-Jones, et al. 2003. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages

256–263, Prague, Czech Republic. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

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, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Jacob Eisenstein. 2019. Measuring and modeling language change. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 9–14, Minneapolis, Minnesota. Association for Computational Linguistics.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of science. *Science*, 359(6379).

Aaron Gerow, Yuening Hu, Jordan Boyd-Graber, David M Blei, and James A Evans. 2018. Measuring discursive influence across scholarship. *Proceedings of the national academy of sciences*, 115(13):3308–3313.

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.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International conference on social informatics*, pages 41–57. Springer.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. 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.

Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

H Inhaber and KJSSoS Przednowek. 1976. Quality of research and the nobel prizes. *Social Studies of Science*, 6(1):33–50.

Sarvnaz Karimi, Jie Yin, and Jiri Baum. 2015. Evaluation Methods for Statistically Dependent Text. *Computational Linguistics*, 41(3):539–548.

Bryan T Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2018. Measuring technological innovation over the long run. *NBER Working Paper*, (w25266).

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.

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*, pages 625–635. International World Wide Web Conferences Steering Committee.

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.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Stephen Lawani. 1986. Some bibliometric correlates of quality in scientific research. *Scientometrics*, 9(1-2):13–25.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398, Berlin, Germany. Association for Computational Linguistics.

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.

Alessandro Moschitti. 2004. A study on convolution kernels for shallow statistic parsing. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 335–342, Barcelona, Spain.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. 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 474–484, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word 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 66–76, Hong Kong, China. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Sandeep Soni, Lauren F Klein, and Jacob Eisenstein. 2021a. Abolitionist networks: Modeling language change in nineteenth-century activist newspapers. *Journal of Cultural Analytics*, 6(1):18841.

Sandeep Soni, Kristina Lerman, and Jacob Eisenstein. 2021b. Follow the leader: Documents on the leading edge of semantic change get more citations. *Journal of the Association for Information Science and Technology*, 72(4):478–492.

Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland. Association for Computational Linguistics.

Elizabeth Closs Traugott and Richard B Dasher. 2001. *Regularity in semantic change*, volume 97. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

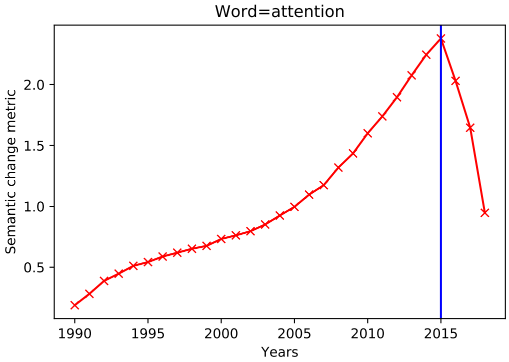

Figure 3: **Visual depiction of change in top example.** Semantic change in the term *attention* in *s2orc*'s ACL anthology subset. The blue line indicates the transition year for meaning change. The transition year for the term *attention* coincides with early papers that described the attention mechanism in neural networks (Bahdanau et al., 2015) that later became the bedrock of transformers architecture (Vaswani et al., 2017)

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Examples of Semantic Changes

We show more statistical details for some of the semantic changes and the context in which these changes occur in Table 5. We also show an illustrative example of a change for the word *attention* and how it transitions according to our metric in Figure 3.

B Topic Coefficients

To control for the content of the paper, we use a coarse-grained representation of the content by learning an LDA model and estimating the probability distribution of a research paper in terms of the topics. The probabilities are used as features in the regression and online prediction tasks. The regression coefficients of the topics in the full model, M4, are shown in Table 6.

C Regression Results for Different Bandwidths

Different lexical and semantic influence estimates were learned by varying the bandwidth (γ). The bandwidth is a decay factor for the influence: higher bandwidth value corresponds to faster decay in influence and a lower bandwidth means a slower decay. The regressions were run for different values of the bandwidth setting $\{.001, .01, .1, 1.0, 10.0, 100.0\}$ and the optimal bandwidth was selected based on the goodness of fit on a 10% heldout sample. The regression results for all the bandwidths are presented in Table 7, Table 8, Table 9, Table 10 and Table 11.

Term	Year	Score	Relative count pre-transition	Relative count post-transition	Earlier usages	Later usages
attention	2015	2.38	126	1670	*increased **attention** over the past several years* *need to be paid **attention***	*parallelizable **attention** networks* *vector of **attention** weights*
plan	2001	1.52	381	158	***plan** such a message* *embedded in the **plan** library*	***plan** recognition problems* ***plan** for tag generation*
network	2013	1.19	240	1000	*semantic **network** path schemata* ***network** of semantically related noun senses*	*deep learning **network** configurations* ***network** parameters to tune*
focus	2006	0.99	451	521	*tracking local **focus*** ***focus** of attention in discourse*	*main **focus** of our work* *the **focus** particle*
representations	2013	0.94	257	1018	*grammatical **representations*** *logical semantic **representations***	*learning distributed **representations*** *learned **representations** across views*
deep	2014	0.94	114	417	***deep** cognitive understanding* ***deep** syntactic features*	***deep** learning* ***deep** architectures*

Table 5: **Semantic change examples.** Top examples of semantic changes identified from the curated ACL corpus from the *s2orc* dataset. The relative counts are counts per million tokens. Terms such as *attention* get a new sense increasingly used later; terms such as *plan* shows semantic widening moving from strong association with dialogue to other NLP tasks; terms such as *network* and *deep* show semantic narrowing moving from disperse associations to a more narrower sense associated with neural networks.

Topic	Regression coefficients	Top words by probability
5	-0.217	*user*(0.017), *users*(0.007), *speech*(0.005), *knowledge*(0.005), *generation*(0.004)
6	-0.230	*query*(0.016), *similarity*(0.015), *term*(0.014), *documents*(0.012), *candidate*(0.011)
13	-0.064	*dialogue*(0.030), *domain*(0.022), *utterance*(0.013), *dialog*(0.012), *utterances*(0.012)
15	-0.064	*image*(0.024), *visual*(0.020), *object*(0.017), *objects*(0.016), *spatial*(0.013)
14	-0.043	*translation*(0.057), *source* (0.028), *target*(0.022), *alignment*(0.018), *parallel*(0.013)
12	-0.065	*speech*(0.015), *chinese*(0.015), *character*(0.013), *languages*(0.013), *segmentation*(0.009)
10	-0.005	*lexical*(0.010), *syntactic*(0.010), *verbs*(0.008), *noun*(0.007), *argument*(0.007)
20	0	*tree*(0.021), *node*(0.018), *nodes*(0.013), *rule*(0.011), *rules* (0.011)
2	0.011	*classification*(0.021), *classifier* (0.020), *class*(0.013), *discourse*(0.011), *accuracy*(0.010)
17	0.122	*dependency*(0.033), *parsing* (0.027), *parser*(0.023), *syntactic*(0.019), *parse*(0.015)
3	0.085	*event*(0.041), *annotation*(0.031), *events*(0.020), *coreference*(0.018), *mentions*(0.013)
9	0.117	*morphological*(0.018), *pos*(0.017), *tag*(0.012), *tags*(0.011), *languages*(0.009)
11	0.106	*question*(0.029), *answer*(0.024), *questions*(0.019), *attention*(0.013), *dataset*(0.012)
18	0.143	*sentiment*(0.027), *tweets*(0.013), *negative*(0.011), *positive*(0.011), *opinion*(0.011)
16	0.149	*sense*(0.023), *similarity*(0.015), *wordnet*(0.012), *senses*(0.011), *target* (0.009)
19	0.122	*topic*(0.032), *document*(0.026), *summary*(0.014), *documents*(0.014), *topics*(0.013)
7	0.167	*human*(0.006), *texts*(0.005), *study*(0.005), *had*(0.004), *linguistic*(0.004)
4	0.174	*relation*(0.036), *entity*(0.033), *relations*(0.025), *entities*(0.021), *knowledge*(0.016)
1	0.226	*probability*(0.016), *algorithm* (0.010), *distribution*(0.009), *parameters* (0.007), *function*(0.006)
8	0.334	*neural*(0.017), *embeddings*(0.016), *vector*(0.013), *network*(0.012), *embedding*(0.012)

Table 6: **Topic coefficients and top words** We show the coefficients of the topic in the experimental model M4 alongwith the top words by probability in a given topic.

Predictors	M1	M2	M3	M4
Constant	-0.000 (0.005)	-0.080 (0.036)	-0.090 (0.036)	-0.097 (0.037)
Initial Citations	0.763 (0.005)	0.740 (0.005)	0.737 (0.005)	0.731 (0.005)
Lex. Inf. Q2			0.006 (0.011)	0.004 (0.011)
Lex. Inf. Q3			0.020 (0.014)	0.015 (0.014)
Lex. Inf. Q4			0.066 (0.016)	0.054 (0.016)
Sem. Inf. Q2				-0.007 (0.011)
Sem. Inf. Q3				0.018 (0.014)
Sem. Inf. Q4				0.135 (0.017)
Log Lik.	-18828	-18681	-18672	-18638

Table 7: **Regression analysis**. We show the results of long-term citations for various ablations when the bandwidth was set to 0.001. The interpretation of columns and rows is similar to Table 3.

Predictors	M1	M2	M3	M4
Constant	-0.000 (0.005)	-0.080 (0.036)	-0.095 (0.036)	-0.102 (0.037)
Initial Citations	0.763 (0.005)	0.740 (0.005)	0.737 (0.005)	0.731 (0.005)
Lex. Inf. Q2			0.010 (0.011)	0.008 (0.011)
Lex. Inf. Q3			0.040 (0.014)	0.036 (0.014)
Lex. Inf. Q4			0.060 (0.016)	0.048 (0.016)
Sem. Inf. Q2				-0.009 (0.011)
Sem. Inf. Q3				0.021 (0.014)
Sem. Inf. Q4				0.133 (0.017)
Log Lik.	-18828	-18681	-18672	-18638

Table 8: **Regression analysis**. We show the results of long-term citations for various ablations when the bandwidth was set to 0.01. The interpretation of columns and rows is similar to Table 3.

Predictors	M1	M2	M3	M4
Constant	-0.000 (0.005)	-0.080 (0.036)	-0.103 (0.036)	-0.115 (0.037)
Initial Citations	0.763 (0.005)	0.740 (0.005)	0.736 (0.005)	0.729 (0.005)
Lex. Inf. Q2			0.026 (0.011)	0.022 (0.011)
Lex. Inf. Q3			0.036 (0.014)	0.028 (0.014)
Lex. Inf. Q4			0.094 (0.016)	0.078 (0.016)
Sem. Inf. Q2				0.008 (0.011)
Sem. Inf. Q3				0.036 (0.014)
Sem. Inf. Q4				0.148 (0.017)
Log Lik.	-18828	-18681	-18663	-18625

Table 9: **Regression analysis**. We show the results of long-term citations for various ablations when the bandwidth was set to 0.1. The interpretation of columns and rows is similar to Table 3.

Predictors	M1	M2	M3	M4
Constant	-0.000 (0.005)	-0.080 (0.036)	-0.085 (0.036)	-0.106 (0.036)
Initial Citations	0.763 (0.005)	0.740 (0.005)	0.736 (0.005)	0.723 (0.005)
Lex. Inf. Q2			0.032 (0.012)	0.021 (0.012)
Lex. Inf. Q3			0.047 (0.014)	0.029 (0.014)
Lex. Inf. Q4			0.092 (0.017)	0.063 (0.017)
Sem. Inf. Q2				0.049 (0.012)
Sem. Inf. Q3				0.097 (0.015)
Sem. Inf. Q4				0.188 (0.018)
Log Lik.	-18828	-18681	-18664	-18603

Table 10: **Regression analysis**. We show the results of long-term citations for various ablations when the bandwidth was set to 1.0. The interpretation of columns and rows is similar to Table 3.

Predictors	M1	M2	M3	M4
Constant	-0.000 (0.005)	-0.080 (0.036)	-0.108 (0.036)	-0.117 (0.036)
Initial Citations	0.763 (0.005)	0.740 (0.005)	0.727 (0.005)	0.718 (0.005)
Lex. Inf. Q2			0.079 (0.012)	0.067 (0.012)
Lex. Inf. Q3			0.097 (0.014)	0.075 (0.014)
Lex. Inf. Q4			0.177 (0.017)	0.141 (0.018)
Sem. Inf. Q2				0.025 (0.012)
Sem. Inf. Q3				0.092 (0.015)
Sem. Inf. Q4				0.157 (0.018)
Log Lik.	-18828	-18681	-18615	-18569

Table 11: **Regression analysis**. We show the results of long-term citations for various ablations when the bandwidth was set to 10.0. The interpretation of columns and rows is similar to Table 3.