Diverging Divergences: Examining Variants of Jensen Shannon Divergence for Corpus Comparison Tasks

Jinghui Lu, Maeve Henchion, Brian Mac Namee
Insight Centre for Data Analytics, Teagasc Agriculture and Food Development Authority
Dublin Ireland, Dublin Ireland
Jinghui.Lu@ucdconnect.ie, Maeve.henchion@teagasc.ie, Brian.MacNamee@ucd.ie

Abstract

Jensen-Shannon divergence (JSD) is a distribution similarity measurement widely used in natural language processing. In corpus comparison tasks, where keywords are extracted to reveal the divergence between different corpora (for example, social media posts from proponents of different views on a political issue), two variants of JSD have emerged in the literature. One of these uses a weighting based on the relative sizes of the corpora being compared. In this paper we argue that this weighting is unnecessary and, in fact, can lead to misleading results. We recommend that this weighted version is not used. We base this recommendation on an analysis of the JSD variants and experiments showing how they impact corpus comparison results as the relative sizes of the corpora being compared change.

Keywords: Corpus Comparison, Jensen-Shannon Divergence

1. Introduction

Corpus comparison approaches are widely used in the digital humanities to identify the differences between corpora through identification of divergent keywords. A symmetric divergence metric, *Jensen-Shannon divergence* (JSD) (Lin, 1991), is one of the most frequently applied techniques used to identify divergent keywords in this type of task. For example, JSD has been used to compare social media posts from different social groups (Gallagher et al., 2016; Lu et al., 2017; Mangold, 2016) or articles from pairs of different years (Pechenick, 2015; Koplenig, 2015; Gerlach et al., 2016; Woodward III, 2016).

When this approach to corpus comparison is used, divergence scores for each word in the two corpora being compared are calculated by JSD and the k words with the highest divergence scores are usually shown to a user, along with the corpus within which they are most prevalent. Visualisations such as those shown in Figure 1 are often used for this. In each back-to-back bar chart, we list the top 30 most divergent terms found within two corpora. Higher ranks indicate more distinguishing words, and the directions of the bars indicate the corpus within which a word is more prevalent. For example, in Figure 1(c) which shows the results of a corpus comparison between articles from two different sections of the New York Times, a bar to the left indicates a term is more common in the *education* section while a bar to the right suggests that a term is more common in the *sports* section.

Two versions of JSD for performing corpus comparison have emerged in the literature (Pechenick et al., 2015a; Gallagher et al., 2016). In this paper we argue, however, that these two variants give dramatically different, and in some cases misleading results, and so more care should be taken in choosing which one to use. In the next section we describe these differences in detail. In Section 3. we demonstrate the impact of using the two variants through a series of experiments performed on large text corpora. Finally, in Section 4. we make recommendations on which variant should be used.

2. Examining JSD

Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) is a statistical measure for estimating the difference between two probability distributions. In natural language processing, a corpus (or a single document) can be regarded as a probability distribution across words in a vocabulary, and the KL divergence between two corpora P and Q can be calculated as:

$$D_{KL}(P\|Q) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \tag{1}$$

where n is the number of unique words in the two corpora; and p_i and q_i are the probabilities of observing word i in corpus P or Q respectively, which is usually estimated through dividing ith word occurrence frequency by the total number of words of the corpus.

As a measure of similarity, KL divergence has the disadvantage of being asymmetric as $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$. Also, when used to compare corpora, a word that only appears in one corpus can result in an infinite KL divergence value.

Lin (1991) proposed Jensen-Shannon divergence (JSD) which is a symmetric version of KL divergence calculated as:

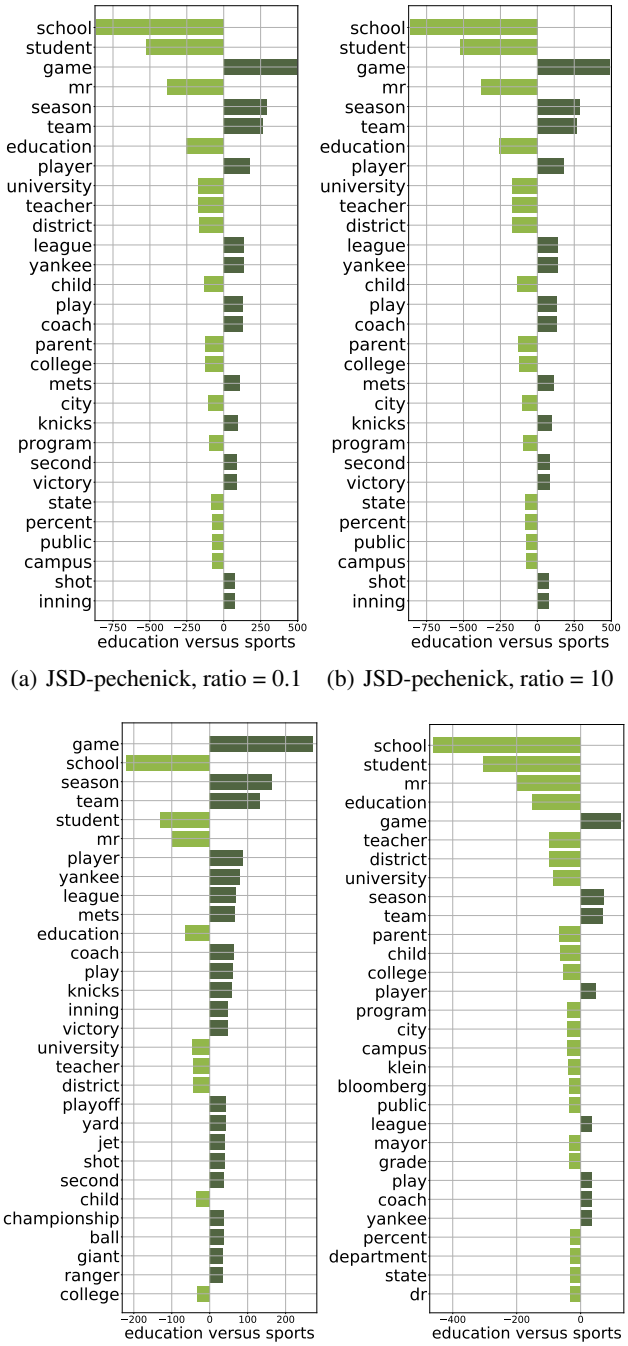
$$D_{JS}(P\|Q) = \frac{1}{2}(D_{KL}(P\|M) + D_{KL}(Q\|M)) \tag{2}$$

where $M = \frac{1}{2}(P+Q)$ is a mixed distribution.

By decomposing Equation 2, the divergence contribution of individual elements in the distribution (words in the case of a corpus) can be computed as:

$$\begin{aligned} D_{JS,i}\left(P \,||\, Q\right) = &-m_i \log_2 m_i \\ &+ \frac{1}{2}(p_i \log_2 p_i + q_i \log_2 q_i) \end{aligned} \tag{3}$$

where m_i is the probability of seeing element i in mixed distribution M.



(a) JSD-pechenick, ratio = 0.1 (b) JSD-pechenick, ratio = 10

(c) JSD-gallagher, ratio = 0.1 (d) JSD-gallagher, ratio = 10

Figure 1: Results of corpus comparisons between corpora of articles from the 1999 *sports* and 2003 *education* sections of the New York Times. The 30 most divergent words are extracted in each case. The words are ranked according to their JSD scores indicated by the bar lengths. The bar to left suggests that a term is more common in the *education* section while the bar to right indicates a term is more common in the *sports* section. Corpus pairs when sports articles out-number education articles at a ratio of 1:10 (Figure 1(a) and 1(c)) and vice versa (Figure 1(b) and 1(d)) have been generated. The corpus comparisons have been repeated using *JSD-pechenick* and *JSD-gallagher*.

Pechenick et al. (2015a; 2015b) were the first to apply JSD to corpus comparison. They examined the evolution of language use in the Google Books corpus (Michel et al., 2011) by computing the JSD contributions of words in corpora of documents from pairs of different years. Since then the method has also been adopted by Koplenig et al. (2015; 2019), Woodward III. (2016) and Gerlach et al. (2016). Pechenick et al. defined JSD using Equation 2 and 3, which we refer to as *JSD-pechenick*.

A variant of JSD later emerged in Gallagher et al. (2016) and has since been widely adopted (e.g. Lu et al. (2017), Mangold (2016), Dennis (2019), Gallagher (2017)). This variant, which we refer to as *JSD-gallagher*, is defined as:

$$\begin{aligned} D_{JS}(P\|Q) = \pi_1 D_{KL}(P\|M_g) \\ +\pi_2 D_{KL}(Q\|M_g) \end{aligned} \tag{4}$$

where $M_g = \pi_1 P + \pi_2 Q$ is a mixed distribution of corpus P and Q; and π_1 and π_2 are weights proportional to the sizes of P and Q, with $\pi_1 + \pi_2 = 1$. Here, the size of a corpus is the total number of words in that corpus. Equation 3, which describes the JSD contributions of individual words, can also be reformulated as:

$$\begin{aligned} D_{JS,i}\left(P \,||\, Q\right) = -m_i \log_2 m_i \\ + \pi_1 p_i \log_2 p_i + \pi_2 q_i \log_2 q_i \end{aligned} \tag{5}$$

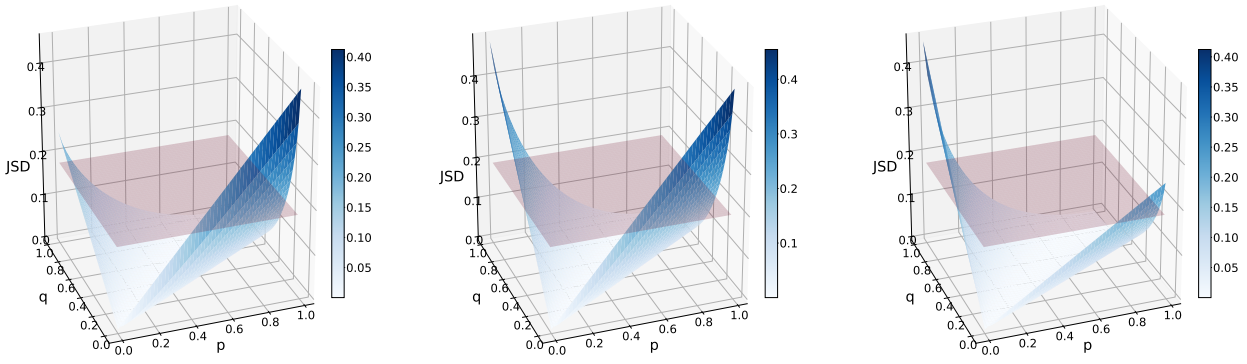
We argue that the corpus size based weighting used in *JSD-gallagher* is unnecessary, and can even undermine the results of corpus comparisons. When the corpora being compared are of different sizes the *JSD-gallagher* variant systematically, and inappropriately, gives higher divergence scores to words that are frequent in the smaller corpus.

Figure 1 shows examples in which corpora of articles from the New York Times *sports* section (throughout 1999) and *education* section (throughout 2003) have been compared. In the corpora used to generate the results in Figure 1(c) the *sports* corpus has been under-sampled to make it is 9 times smaller than the *education* corpus. In this case 21 of the top 30 most divergent terms found in a corpus comparison performed using *JSD-gallagher* are more prevalent in the *sports* section, the smaller corpus. On the other hand, in Figure 1(d) where the data has been sampled to make the *sports* corpus 9 times larger than *education* corpus the pattern is reversed. In contrast, Figures 1(a) and 1(b) show that when JSD-pechenick is used the results are the same regardless of the relative corpus sizes.

The origin of this behaviour can be seen from the shape of the function that *JSD-gallagher* implements. As π_1 and π_2 sum to 1 we can rewrite them as π and $1-\pi$ and refactor Equation 5 as:

$$\begin{aligned} JSD(p,q) = \\ -(\pi p + (1-\pi)q)\log_2(\pi p + (1-\pi)q) \\ +\pi p \log_2 p + (1-\pi)q \log_2 q \\ s.t. \quad 0 < \pi < 1, 0 \leq p \leq 1, 0 \leq q \leq 1 \end{aligned} \tag{6}$$

where p and q are the normalised occurrences of a word in corpus P and Q respectively; π is the weight for corpus P calculated as $\frac{size(P)}{size(P)+size(Q)}$; and $1-\pi$ is the weight for



(a) $\pi = 0.2$ (b) $\pi = 0.5$ (c) $\pi = 0.8$

Figure 2: Plots of $JSD(p, q)$ for different values of π where $\pi = \frac{size(P)}{size(P)+size(Q)}$. The p-axis and q-axis are the relative word frequencies under corpus P, Q respectively. The vertical axis represents the contributions of words computed by *JSD-gallagher* (see Equation 6). Specifically, *JSD-gallagher* is identical to *JSD-pechenick* when $\pi = 0.5$ (see Figure 2(b))

corpus Q. Plots of the surface defined by Equation 6, given different values of π, are shown in Figure 2.

Figure 2(b) shows the surface generated by $JSD(p, q)$ when the corpora are the same size, $\pi = 0.5$ (in this case *JSD-gallagher* is identical to *JSD-pechenick*). The surface is a reverse saddle with a 0 value wherever $p = q$ (representing a word that has the same normalised frequency in both corpora and contributes nothing to divergence). The function has two maximum points at $p = 1, q = 0$ and $p = 0, q = 1$. This represents words that only appear in one corpus and contribute strongly to divergence. In this scenario when extracting the top k divergent words, words from either corpus are equally likely to be selected (illustrated as points above the transparent horizontal plane).

In Figure 2(a) $\pi = 0.2$, and as corpus P is much smaller than corpus Q, a different scenario arises. It is more likely that words from corpus P with very high JSD scores close to $JSD(1, 0)$ will be selected as most divergent. It will be very rare for a word from corpus Q to be selected as most divergent, not because the words do not contribute to divergence, but rather because they come from the larger corpus. Conversely, when $\pi = 0.8$, indicating that corpus P is much larger than Q, *JSD-gallagher* tends to select many more words from corpus Q as highly divergent (see Figure 2(c)).

3. Experiments

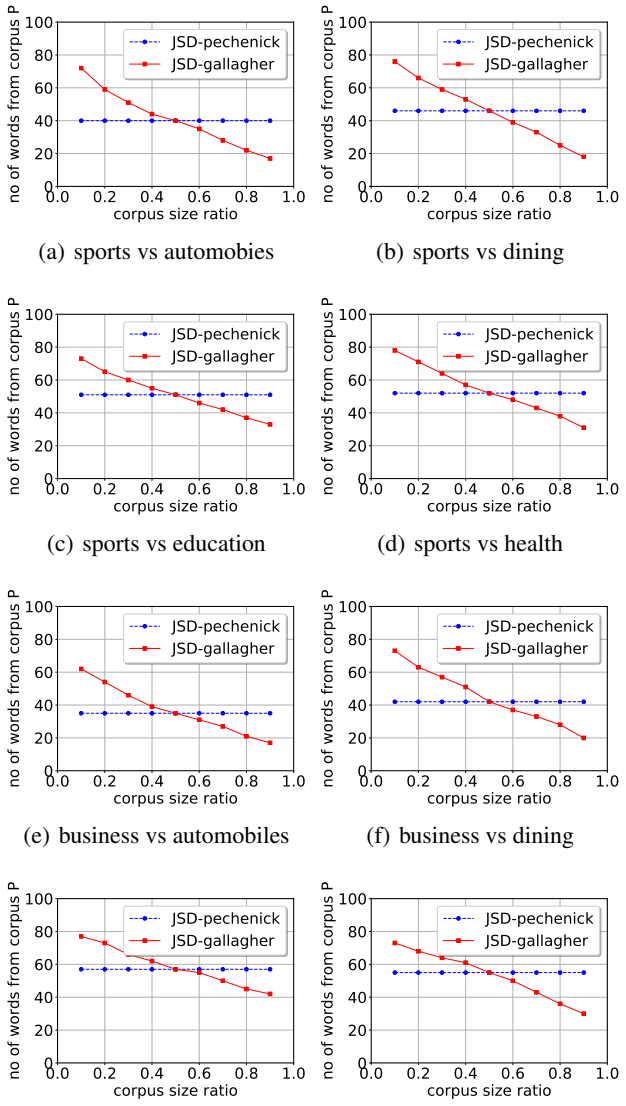
This section describes two experiments demonstrating the behaviour of the JSD variants. In the first we calculate the frequencies of each word in the two corpora independently and introduce imbalance by simply scaling the word frequencies in one corpus. This simple approach isolates the impact of the JSD variants as, although the overall size of the manipulated corpus changes, the relative frequencies of the words within that corpus remain the same, as does the vocabulary used.

This experiment uses corpora from different sections of the New York Times (Greene et al., 2014): *sports*, *business*, *automobies*, *health*, *dining*, and *education*. In the experiment the size of one corpus is left unchanged while the size of the *manipulated corpus* is decreased so that $\frac{size(P)}{size(P)+size(Q)}$ varies from 0.1 to 0.9 in steps of 0.1. This is achieved by scaling calculated word frequencies for the manipulated corpus. Figure 3 shows how the number of words from the unchanged corpus that appear in the list of the top 100 most divergent words changes as the size of the manipulated corpus is decreased. The results shown in Figure 1 also come from this experiment.

Figure 3 shows that when *JSD-gallagher* is used, as the size of the manipulated corpus, Q, decreases fewer words from the unchanged corpus, P, appear in the top 100 most divergent words list. This verifies our claim that *JSD-gallagher* tends to select more words from the smaller corpus in a comparison. The results also show, however, that when *JSD-pechenick* is used the number of words selected as most divergent from each corpus is not affected by relative corpus size.

We also perform a second, more realistic experiment, where imbalance is induced by removing documents from one corpus. As well as the New York Times datasets we also use 6 other datasets (Greene et al., 2014) from Wikipedia (wikipedia-high, wikipedia-low), the BBC (bbc, bbc-sport), The Guardian (guardian-2013), and The Irish Times (irishtimes-2013). We construct 4 corpus pairs, shown in Figure 4. Each pair is constructed to be balanced and then the size of one corpus, Q, is reduced by randomly removing documents from it from 0% to 90% in 10% steps. At each step a corpus comparison is performed and the number of words selected as most divergent from the unchanged corpus is recorded.

Figure 4 shows how the number of words selected as most divergent from the unchanged corpus changes as corpus imbalance increases. All results are averages over 10 independent runs. The number of words selected as most divergent from the unchanged corpus by *JSD-pechenick* changes little as the size of the corpus is changed. The small changes present, are most likely due to changes in vocabulary as



(a) sports vs automobies (b) sports vs dining

(c) sports vs education (d) sports vs health

(e) business vs automobiles (f) business vs dining

(g) business vs education (h) business vs health

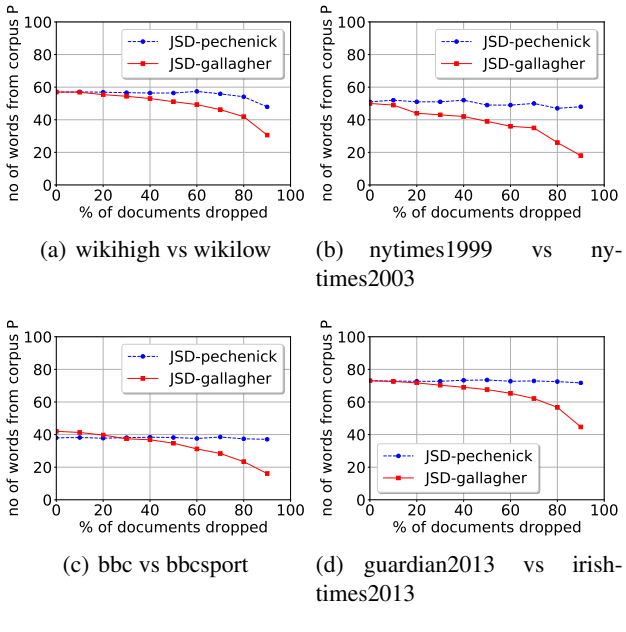
Figure 3: The number of words selected as most divergent by JSD variants as relative corpus size changes (size of corpus Q is decreased by scaling word frequencies). X-axis denotes the corpus size ratio $\frac{size(P)}{size(P)+size(Q)}$, Y-axis represents the number of terms from the manipulated corpus, Q, that appear in the list of top 100 most divergent words.

large numbers of documents are removed. However, the number of words selected as most divergent from the unchanged corpus by *JSD-gallagher* changes dramatically, and the terms from the larger corpus, P, are not favoured as imbalance grows. The initial corpora are constructed to be balanced so at the beginning of each graph *JSD-pechenick* and *JSD-gallagher* behave similarly.

4. Conclusions

In this paper, we have demonstrated that two different variants of JSD that are both commonly used in corpus comparison studies in the literature can lead to very different results. This arises from the weights based on corpus size that are introduced in the *JSD-gallagher* variant that are not used in the *JSD-pechenick* variant. In particular the *JSD-gallagher* variant is especially sensitive to size differences between the corpora being compared. This is a significant issue in comparisons performed between corpora of very different sizes—for example printed media versus online media (Zhao et al., 2011). In that case, most of the top k diverging words will be selected from the smaller corpus if *JSD-gallagher* is used. This is a misleading result as it is only the fact that they are in the smaller corpus which leads to those words being selected, not their divergent properties. We have demonstrated this problem in two experiments using real datasets, as well as demonstrating where in the formulation of the JSD calculation that this effect comes from.

From our analysis we recommend that the *JSD-pechenick* variant of JSD is always used. This version is invariant to changes in relative corpus size and tends towards selections of most divergent words that are balanced between the corpora being compared.



(a) wikihigh vs wikilow (b) nytimes1999 vs nytimes2003

(c) bbc vs bbcsport (d) guardian2013 vs irishtimes2013

Figure 4: The number of words selected as most divergent by JSD variants as relative corpus size changes (documents in corpus Q are randomly removed). X-axis denotes the percentages documents removed from Q and Y-axis represents the number of terms from the unchanged corpus, P, that appear in the list of top 100 most divergent words.

5. Acknowledgements

This research was kindly supported by a Teagasc Walshe Fellowship award (2016053) and Science Foundation Ireland (12/RC/2289_P2).

6. Bibliographical References

Dennis, J. (2019). # stopslacktivism: Why clicks, likes, and shares matter. In *Beyond Slacktivism*, pages 25–69. Springer.

Gallagher, R. J., Reagan, A. J., Danforth, C. M., and Sheridan Dodds, P. (2016). Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. *arXiv preprint arXiv:1606.06820*.

Gallagher, R. (2017). Disentangling discourse: Networks, entropy, and social movements.

Gerlach, M., Font-Clos, F., and Altmann, E. G. (2016). Similarity of symbol frequency distributions with heavy tails. *Physical Review X*, 6(2):021009.

Greene, D., O'Callaghan, D., and Cunningham, P. (2014). How many topics? stability analysis for topic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 498–513. Springer.

Koplenig, A., Wolfer, S., and Müller-Spitzer, C. (2019). Studying lexical dynamics and language change via generalized entropies: The problem of sample size. *Entropy*, 21(5):464.

Koplenig, A. (2015). A fully data-driven method to identify (correlated) changes in diachronic corpora. *arXiv preprint arXiv:1508.06374*.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Lu, J., Henchion, M., and Namee, B. M. (2017). Extending jensen shannon divergence to compare multiple corpora. In *AICS*.

Mangold, L. (2016). should i stay or should i go: Clash of opinions in the brexit twitter debate. *Computing*, 1(4.1).

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015a). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015b). Is language evolution grinding to a halt?: Exploring the life and death of words in english fiction. *arXiv preprint arXiv:1503.03512*.

Pechenick, E. (2015). Exploring the google books corpus: An information-theoretic approach to linguistic evolution.

Woodward III, R. B. (2016). Quantifying cultural changes through a half-century of song lyrics and books.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338–349. Springer.